Pedagogia Jurnal Ilmu Pendidikan

Journal homepage: http://ejournal.upi.edu/index.php/pedagogia

Maximizing Learning Outcomes: A Comparative Analysis of IRT- and CTT- Differentiated Learning based Design

*Dwi Rismi Ocy, Awaluddin Tjalla, Soeprijanto

*Department of Educational Research and Evaluation, Universitas Negeri Jakarta, Jakarta, Indonesia

*Correspondence: E-mail: dwirismiocy@gmail.com

ABSTRACT

Challenges in addressing diverse student abilities often hinder effective learning, particularly in complex subjects like linear equations and inequalities. This research aimed to compare the effectiveness of Item Response Theory (IRT)-based and Classical Test Theory (CTT)-based differentiated learning designs in improving student performance in linear equations and inequalities. Conducted in two secondary schools, the study involved 126 students, with 61 students in the IRT group and 65 students in the CTT group. A quasi-experimental design with pretest-posttest control groups was employed to assess learning progress. The results showed that both IRT and CTT-based learning interventions led to significant improvements in student performance. However, the IRT-based approach, which grouped students based on their individual ability levels and tailored tasks to their proficiency, resulted in a significantly higher average posttest score and a very large effect size. The CTT-based approach also showed improvement but with a smaller effect size. The findings suggest that IRT offers a more precise and effective method for differentiating instruction, leading to better learning outcomes, particularly in complex subjects like linear equations. This study underscores the potential of IRT in enhancing educational practices and improving student learning outcomes.

ARTICLE INFO

Article History:
Submitted/Received 19 Aug 2024
First Revised 19 Oct 2024
Accepted 17 Nov 2024
First Available Online 28 Nov 2024
Publication Date 01 Dec 2024

Keyword:
Differentiated Instruction, Pedagogical Strategies, Student-Centered Learning, Item Response Theory (IRT), Classical Test Theory (CTT).

1. INTRODUCTION

The choice of learning approach plays a critical role in shaping student success, especially in acquiring foundational mathematical concepts such as linear equations and inequalities. Linear equations are fundamental concepts in mathematics that lay the groundwork for more advanced topics in algebra, making them essential in fostering students' mathematical proficiency. As educators strive to optimize learning, assessment methods are integral in guiding instructional strategies. Traditional approaches, such as Classical Test Theory (CTT), have long dominated educational assessments, offering insights through basic measures like total scores. However, with the advancement of Item Response Theory (IRT), a more refined approach has emerged. IRT offers a deeper understanding of student abilities and item difficulty, providing more granular insights into the learning process (Cardamone et al., 2012).

Item Response Theory, by estimating both student ability and item difficulty based on response patterns, enhances the educational experience by personalizing learning interventions. IRT models the nonlinear relationship between student abilities and the probability of answering questions correctly, allowing for a more accurate identification of problem areas and appropriate interventions. Unlike CTT, which focuses on total scores, IRT's detailed analysis at the item level offers richer insights into the challenges students face (Ju & Bork, 2005; Van Der Linden & Glas, 2010). This precision empowers educators to design interventions that are better suited to the needs of individual students, optimizing the learning process (Chen et al., 2021).

In contrast, CTT provides a more generalized view of student performance, relying primarily on the count or fraction of correct answers, which may fail to capture the underlying nuances of student ability. This approach can overlook key aspects of student learning and fail to differentiate between students with varying proficiency levels (Dumont & Ready, 2023; Raykov & Marcoulides, 2016). While CTT remains a valuable tool in educational assessment, the limitations it poses in understanding the complexity of student abilities call for alternative methods, such as IRT, to provide a more precise and comprehensive analysis of student performance (Lee, 2019).

Research has shown that IRT provides more accurate estimates of student ability, particularly when assessing complex problems or students with weaker academic skills (Devayanti et al., 2023; Lee, 2019). This suggests that IRT-based assessment can be particularly beneficial in understanding the specific learning needs of students, thereby allowing for the design of more targeted and effective interventions (Raykov & Marcoulides, 2016; Tian et al., 2017). However, a gap in the literature exists regarding direct comparisons between IRT-based and CTT-based learning designs, particularly for subjects like linear equations. This gap presents a valuable opportunity to explore the comparative effectiveness of these two approaches in improving student outcomes.

Previous studies indicate that IRT-based learning designs can significantly enhance student understanding by tailoring instruction to meet individual needs. By analyzing item-level performance data, IRT offers educators a clearer picture of where students are struggling, enabling more effective instructional strategies. In contrast, CTT's more generalized approach may result in less targeted interventions, making it harder to address specific learning gaps (Kubsch et al., 2022; Tetzlaff et al., 2021). The present study, conducted in Indonesian secondary schools, seeks to fill this gap by comparing the effectiveness of IRT and CTT-based learning designs in improving students' ability to solve linear equations and inequalities. This comparison could provide important insights for educators seeking to optimize their assessment and teaching practices (Raykov & Marcoulides, 2016; Tian et al., 2017).

DOI: https://doi.org/10.17509/pdgia.v22i3.77064
p- ISSN p. ISSN 1693-5276 e- ISSN 2579-7700

By exploring how IRT and CTT-based learning designs influence the development and implementation of targeted learning interventions, this research hopes to contribute valuable insights into the design of more effective educational practices. Through a rigorous analysis of both approaches, the study aims to inform future instructional strategies that are not only effective but also personalized to meet the needs of each student (Alsariera et al., 2022; Willis et al., 2022). Ultimately, this study aims to evaluate the comparative effectiveness of IRT and CTT-based learning designs in improving student performance in mathematics, particularly in the area of linear equations.

2. METHODOLOGY

2.1 Research Design

This study employed a quantitative research approach using a quasi-experimental design to compare the effectiveness of Item Response Theory (IRT)-based and Classical Test Theory (CTT)-based learning designs in enhancing students' understanding of linear equations and linear inequalities. A pretest-posttest control group design was utilized to assess the learning gains between two student groups who received different instructional treatments. Group A, consisting of 61 students from SMPIT Al Khoiriyah Al Husna, was taught using an IRT-based learning design. Group B, composed of 65 students from SMPN 1 Sukalarang, received instruction through a CTT-based learning model.

The participants in this study were junior secondary school students enrolled in two different schools. Group A included 61 students, while Group B consisted of 65 students. The schools were selected purposively based on accessibility and their willingness to implement differentiated instructional designs. Both groups participated in the same sequence of instructional content but were exposed to distinct assessment and evaluation frameworks aligned with IRT and CTT principles.

2.2 Data Collection

Data were collected through two primary instruments: a pretest and a posttest. The pretest was administered to evaluate students' baseline understanding and prerequisite knowledge of linear equations and inequalities prior to the learning intervention. Following the instructional period, the posttest was used to measure the extent of students' conceptual understanding and problem-solving skills. The test items were identical for both groups to ensure the comparability of results. For Group A, data from the tests were analyzed using the Rasch Model of IRT to obtain detailed insights into individual student ability and item difficulty. For Group B, data were analyzed using the CTT framework, which relies on aggregate test scores to determine student performance.

2.3 Data Analysis

The effectiveness of the two instructional approaches was assessed using both descriptive and inferential statistics. Independent samples t-tests were used to determine whether there were statistically significant differences between the pretest and posttest results of Group A and Group B. Additionally, Cohen's d effect size formula was employed to evaluate the magnitude of the learning gains resulting from each instructional model. The comparative analysis enabled the researchers to determine not only the statistical significance but also the practical significance of implementing either IRT-based or CTT-based designs in mathematics instruction.

DOI: https://doi.org/10.17509/pdgia.v22i3.77064
p- ISSN p. ISSN 1693-5276 e- ISSN 2579-7700

3. RESULT AND DISCUSSION

3.1 Result

The validity of the test instrument was evaluated using three complementary methods: the Delphi Test, Aiken's V, and the Content Validity Index (I-CVI). These methods collectively ensure the instrument is both theoretically sound and practically applicable. The results of these validations highlight the instrument's strong alignment with its intended purpose, assessing students' understanding and application of linear equations and inequalities.

Table 1. Instrument Test Validity Results

Indicator	Delphi Test (3 Experts)	Aiken's V (21 Teachers/Practitioners)	I-CVI (Content Validity Index)	Remarks
Understanding of Basic Concepts	4.7	0.90	1.0	Valid
Identifying Components	4.7	0.85	1.0	Valid
Solving Linear Equations	5.0	0.92	1.0	Valid
Understanding Inequalities	4.7	0.88	1.0	Valid
Graphing Linear Equations	4.7	0.91	1.0	Valid
Interpreting Solutions	4.7	0.89	1.0	Valid
Solving Linear Inequalities	5.0	0.87	1.0	Valid
Application of Concepts	4.7	0.89	1.0	Valid
Critical Thinking and Reasoning	4.7	0.93	1.0	Valid
Connecting Concepts	4.7	0.91	1.0	Valid

As indicated in Table 1, the validity of the research instrument was confirmed through expert judgment and practitioner evaluation. The Delphi Test, involving three subject-matter experts, produced high average ratings ranging from 4.7 to 5.0 on a 5-point Likert scale, demonstrating strong agreement regarding item clarity and theoretical alignment. Additionally, Aiken's V analysis with 21 teachers and practitioners yielded coefficients between 0.85 and 0.93, exceeding the minimum threshold of 0.75 and affirming the instrument's instructional relevance. Complementing these findings, the Item Content Validity Index (I-CVI) for all items reached 1.0, reflecting unanimous consensus on the content's appropriateness.

Table 2 presents the pretest results for Group A, evaluated using the IRT-Rasch Model. Item difficulty values ranged from -1.73 to 1.71, with a mean person measure of -0.56 and a

DOI: https://doi.org/10.17509/pdgia.v22i3.77064
p- ISSN p. ISSN 1693-5276 e- ISSN 2579-7700

Table 2. Pre-Test Results

Indicator	Group B (CTT) $N = 65$				Group A (IRT - Rasch Model) $N = 61$				
	Difficulty Index	Difficulty Category	Average Score	Reliability (Cronbach Alpha)	Item Measure (Logit Scale)	Standard Deviation (SD)	Difficulty Category	Mean Person Measure (Logit Scale)	Reliability (Cronbach Alpha KR-20)
Understanding of Basic Concepts	0.70	Easy	35.2	0.85	-1.73	1.45	Easy	-0,56	0.82
Identifying Components	0.45	Moderate			-1.25		Moderate		
Solving Linear Equations	0.40	Moderate			-0.48		Moderate		
Understanding Inequalities	0.52	Moderate			0.46		Moderate		
Graphing Linear Equations	0.0	Difficult			1.71		Difficult		
Interpreting Solutions	0.40	Moderate			-0.54		Moderate		
Solving Linear Inequalities	0.39	Difficult			0.18		Moderate		
Application of Concepts	0.10	Difficult			1.62		Difficult		
Critical Thinking and Reasoning	0.05	Difficult			1.09		Moderate		
Connecting Concepts	0.18	Difficult			-0.06		Moderate		

DOI: https://doi.org/10.17509/pdgia.v22i3.77064
p- ISSN p. ISSN 1693-5276 e- ISSN 2579-7700

standard deviation of 1.45, indicating a moderate dispersion in student proficiency. The reliability coefficient of 0.82 suggests acceptable internal consistency. In contrast, Table 3 shows that Group B, assessed through Classical Test Theory (CTT), achieved an average score of 35.2, with item difficulty indices ranging from 0.0 to 0.70. While several items proved challenging, the reliability coefficient of 0.85 indicates strong internal consistency, marginally higher than that of Group A.

In this study, two distinct approaches to Differentiated Learning Design were implemented: (1) the IRT-based Differentiated Learning Design and (2) the CTT-based Differentiated Learning Design. The main objective of this comparison was to explore the effectiveness of these two approaches in addressing the diverse learning needs of students, particularly when grouping students according to their ability levels.

Table 3. Comparison of Learning Designs for IRT-based and CTT-based Groups

Aspect	Group A (IRT)	Group B (CTT)
Grouping Method	Students are grouped based on their ability estimates (θ values) derived from the IRT model. These estimates consider the probability of correctly answering test items with varying difficulty levels, providing a precise measurement of ability.	Students are grouped based on percentile ranks, as calculated using the CTT framework. This approach relies on the total score as a reflection of student ability.
Group Composition	The ability estimates are divided into three levels using cut-off thresholds: High (Logit>SD), Medium (-SD<Logit≤SD), Low (Logit≤SD). These thresholds align with the distribution of the latent trait (θ) and consider item difficulty parameters.	The raw test scores are divided into three categories using statistical percentiles: High (Scores above the 75th Percentile), Medium (Scores between the 25th and 75th Percentiles), Low (Scores below the 25th Percentile).
Differentiated Tasks	Instructional tasks are tailored to the ability levels identified by the IRT model. For example: • High ability: Solve complex, multi-step problems. • Medium ability: Engage in structured problem-solving with some scaffolding. • Low ability: Focus on foundational concepts and practice basic skills.	Instructional tasks are designed similarly to the IRT-based groups but are assigned based on raw score groupings. This may result in less precision, as tasks are matched to broader ability categories.
Assessment Approach	Pre- and post-tests are analyzed using IRT. This enables: • Detection of learning gains at a granular level. • Assessment of changes in ability estimates (θ values).	Pre-and post-tests are analyzed using CTT. This approach measures: • Raw score improvements. • Overall test reliability (Cronbach's alpha)

DOI: https://doi.org/10.17509/pdgia.v22i3.77064
p- ISSN p. ISSN 1693-5276 e- ISSN 2579-7700

Aspect	Group A (IRT)	Group B (CTT)
Flexibility in Adjustments	The IRT model allows dynamic adjustments based on student ability. For instance, high-performing students in the medium group can be reassigned to the high group during instruction, ensuring alignment with their true ability.	Adjustments are less frequent due to the lack of fine-grained diagnostics. Group reassignments are typically based on teacher observation rather than quantitative insights.

As shown in Table 3, Group A applies the IRT model, which categorizes students based on ability estimates (θ values), enabling more accurate differentiation and tailored instruction. Group B uses the CTT model, relying on percentile ranks derived from raw scores, offering a simpler but less diagnostic approach. The IRT framework supports precise grouping and dynamic adjustment of learning tasks, while CTT offers broad classification with limited adaptability. These distinctions result in more targeted interventions for Group A and general instructional strategies for Group B.

Table 4. Group Composition

Category	Person Measure (Logit Scale)	Group A (IRT-Rasch Model)		Group B (CTT)	
		Standard Deviation (SD)	Number of Students	Total Score	Number of Students
High	Logit>SD	1.31	2	Scores above the 75th Percentile	4
Medium	-SD<Logit≤SD		36	Scores between the 25th and 75th Percentiles	33
Low	Logit≤SD		23	Scores below the 25th Percentile	28

As summarized in Table 4, Group A employed the Item Response Theory (IRT) model to classify students into high, medium, and low ability levels based on estimated ability values (θ), enabling differentiated instruction aligned with student proficiency. High-ability learners tackled advanced problem-solving tasks, medium-ability students engaged in guided practice, and low-ability learners received foundational support. In contrast, Group B, using Classical Test Theory (CTT), categorized students by percentile ranks, offering broader but less precise differentiation. While effective, the CTT approach lacks the adaptive precision of IRT in tailoring instructional tasks.

Table 5 presents the post-test outcomes. Group A's IRT-based assessment showed item difficulties from −1.36 to 1.12 and a mean person measure of 0.93, indicating above-average performance. The reliability coefficient of 0.89 affirms strong internal consistency. Group B, assessed through CTT, recorded an average score of 70.5 with difficulty indices from 1.00 to 0.20 and a reliability coefficient of 0.84. Although both models improved learning outcomes, the IRT approach yielded more refined diagnostic insights and better instructional alignment.

DOI: https://doi.org/10.17509/pdgia.v22i3.77064
p- ISSN p. ISSN 1693-5276 e- ISSN 2579-7700

Table 5. Post-Test Results

Indicator	Group B (CTT) $N = 65$				Group A (IRT-Rasch Model) $N = 61$				
	Difficulty Index	Difficulty Category	Average Score	Reliability (Cronbach Alpha)	Item Measure (Logit Scale)	Standard Deviation (SD)	Difficulty Category	Mean Person Measure (Logit Scale)	Reliability (Cronbach Alpha KR-20)
Understanding of Basic Concepts	1.00	Easy	70.5	0.84	-1.36	0.87	Easy	0.93	0.89
Identifying Components	0.98	Easy			-1.00		Easy		
Solving Linear Equations	0.60	Moderate			-0.50		Easy		
Understanding Inequalities	0.69	Moderate			-0.12		Easy		
Graphing Linear Equations	0.27	Difficult			0.22		Moderate		
Interpreting Solutions	0.66	Moderate			0.58		Moderate		
Solving Linear Inequalities	0.54	Moderate			0.68		Moderate		
Application of Concepts	0.35	Difficult			1.12		Difficult		
Critical Thinking and Reasoning	0.20	Difficult			0.35		Moderate		
Connecting Concepts	0.42	Moderate			0.77		Moderate		

DOI: https://doi.org/10.17509/pdgia.v22i3.77064
p- ISSN p. ISSN 1693-5276 e- ISSN 2579-7700

Table 6. Statistical Results of Pretest and Posttest Scores

Statistic	Pretest	Posttest	Interpretation
		Group A (IRT)	
Average Score	34.8	81.6	Scores increased significantly, showing higher effectiveness of IRT design.
Normality (Shapiro-Wilk)	W = 0.980,p = 0.154	W = 0.970,p = 0.095	Data are normally distributed (p > 0.05), satisfying parametric test assumptions.
Homogeneity (Levene's)	F(1,124) = 0.845,p = 0.361	F(1,124) = 1.021,p = 0.315	Variances are equal between groups (p > 0.05), validating the t-test.
Paired t-test	t(60) = 18.43,p < 0.001	—	Significant improvement from pretest to posttest (p < 0.001).
		Group B (CTT)	
Average Score	35.2	70.5	Scores increased significantly, showing improved performance after intervention.
Normality (Shapiro-Wilk)	W = 0.972,p = 0.123	W = 0.963,p = 0.085	Data are normally distributed (p > 0.05), satisfying parametric test assumptions.
Homogeneity (Levene's)	F(1,124) = 0.845,p = 0.361	F(1,124)= 1.021,p=0.315	Variances are equal between groups (p > 0.05), validating the t-test.
Paired t-test	t(64) = 14.12,p < 0.001	—	Significant improvement from pretest to posttest (p < 0.001).
Independent t-test	t(124)= -0.08,p = 0.936	t(124) = -4.87,p < 0.001	Pretest: No significant difference (p > 0.05); Posttest: Significant difference (p < 0.001).

As shown in Table 6, statistical analyses were conducted to compare the effectiveness of IRT- and CTT-based differentiated learning. Shapiro-Wilk and Levene's tests confirmed that the data met assumptions of normality and homogeneity of variance. An independent t-test revealed no significant difference in pretest scores between Group A (IRT) and Group B (CTT), indicating comparable baseline knowledge. However, posttest results showed a significant difference favoring Group A, with a higher average score. Paired t-tests for both groups indicated significant improvement from pretest to posttest.

Table 7. Effect Size Comparison of IRT-and CTT-Based Differentiated Learning Design

Group	Pretest Average Score	Posttest Average Score	Effect Size (Cohen's d)	Cohen's d Category
Group A (IRT-Rasch Model)	34.8	81.6	4.25	Very Large
Group B (CTT)	35.2	70.5	3.20	Large

DOI: https://doi.org/10.17509/pdgia.v22i3.77064
p- ISSN p. ISSN 1693-5276 e- ISSN 2579-7700

As shown in Table 7, both groups demonstrated significant gains in learning outcomes. Group A's average score increased from 34.8 to 81.6, yielding a very large effect size (Cohen's d = 4.25), indicating a highly effective impact of the IRT-based intervention. Group B improved from 35.2 to 70.5, with a large effect size (Cohen's d = 3.20), confirming the CTT approach was also effective, though less pronounced. These findings highlight the superior impact of IRT-based differentiated learning on student performance in linear equations and inequalities.

3.2 Discussion

The findings of this study underscore the relative strengths of Item Response Theory (IRT)-based and Classical Test Theory (CTT)-based differentiated learning designs in enhancing student comprehension of linear equations and inequalities. Both approaches yielded measurable improvements in student outcomes, but the IRT-based model demonstrated a greater capacity for individualized instructional support and learning effectiveness (Gilbert et al., 2023). Instrument validity was established through a triangulated validation process involving expert review and practitioner evaluation. High consensus on the clarity, relevance, and alignment of test items affirms the robustness of the instrument used to assess learning outcomes. These strong validation outcomes lend credibility to the data interpretations and suggest that the assessments were both conceptually rigorous and educationally appropriate (Cook et al., 2016; Cook & Hatala, 2016).

Pretest outcomes revealed that students displayed varying levels of initial understanding, highlighting the need for differentiated instructional strategies. The IRT-based model provided a more detailed profile of student competencies by accounting for both individual performance and item difficulty. This contrasted with the CTT model, which offered a general snapshot based on total scores, offering less insight into specific learning needs. While both models produced reliable assessments, the IRT approach allowed for greater diagnostic precision, supporting more tailored instructional decisions (Csapó & Molnár, 2019; Larrain & Kaiser, 2022; Pliakos et al., 2019).

The design differences between the two instructional models had important pedagogical implications. The IRT model enabled dynamic grouping of students based on ability estimates, which facilitated more accurate alignment between student needs and instructional content. In contrast, the CTT model relied on static percentile rankings, making it more limited in responsiveness. The adaptability of the IRT framework allowed educators to continuously recalibrate instructional strategies, thereby enhancing the relevance and effectiveness of learning interventions (Mallillin, 2022; Pak et al., 2022).

When translated into classroom practices, the impact of these models became more pronounced. Students in the IRT-based group received instructional tasks carefully aligned with their demonstrated abilities, ranging from foundational skill development to complex problem-solving and reasoning. The CTT-based group also experienced differentiated learning, but with broader categorizations that may have obscured more nuanced learning needs. As a result, instructional alignment was more precise and impactful in the IRT context (Becker & Nekrasova-Beker, 2018; Diaz et al., 2023; Eren et al., 2023).

Post-instruction assessments further affirmed the effectiveness of the IRT approach. Students showed strong mastery of foundational and intermediate concepts and demonstrated growth in higher-order thinking tasks. The ability of the IRT model to align instruction closely with individual proficiency levels contributed to this improved performance. Meanwhile, the CTT-based group also exhibited learning gains, particularly on foundational tasks, though progress on more complex items was more limited. This suggests

DOI: https://doi.org/10.17509/pdgia.v22i3.77064
p- ISSN p. ISSN 1693-5276 e- ISSN 2579-7700

that while both models foster learning, the precision of the IRT model offers added instructional value (Abedalaziz & Leng, 2018; Carlson & von Davier, 2017).

The results of statistical analyses reinforce the superiority of the IRT-based design. Both instructional models were effective in improving student outcomes, confirming the general utility of differentiated instruction. However, only the IRT-based group showed a clearly distinguishable advantage in post-instruction performance, suggesting that this model is better suited to address diverse learning profiles within a classroom (Bustamante & Navarro, 2022). Effectiveness comparisons further illustrate that the impact of IRT-based instruction extended beyond surface-level score gains. The instructional responsiveness enabled by continuous ability estimation allowed for deeper student engagement and better alignment between tasks and learner readiness. The CTT-based approach, although simpler to implement, lacked this adaptive capacity, making it less responsive to evolving student needs (Beggrow et al., 2014; Tornabene et al., 2018).

These findings support earlier work advocating for psychometric-informed instruction. The IRT model's ability to inform nuanced instructional differentiation makes it especially promising for complex domains like mathematics, where student abilities often span a wide spectrum. It provides a mechanism not only for categorizing student performance but also for strategically responding to individual learning needs (Hitt & Tucker, 2016; Pea, 2018).

4. CONCLUSION

This study compared the effectiveness of two differentiated learning designs: an IRT-based approach (Group A) and a CTT-based approach (Group B) in enhancing student performance on linear equations and inequalities. The results of the statistical analyses showed that both approaches significantly improved student learning outcomes, as evidenced by the significant increase in posttest scores for both groups. However, the IRT-based approach proved to be more effective in promoting student performance. Group A, which utilized the IRT model for differentiating instruction based on individual ability levels, showed a significantly higher posttest average score and a very large effect size compared to Group B, which used the CTT approach.

The IRT-based approach allowed for more precise grouping of students based on their estimated ability levels, enabling more targeted and individualized instructional strategies. This resulted in greater learning gains and a more personalized learning experience for the students. In contrast, the CTT-based approach, while still effective, was less precise in categorizing students and assigning differentiated tasks, resulting in a more generalized improvement in learning outcomes.

The findings suggest that incorporating IRT into the design of differentiated learning can offer significant advantages in terms of understanding student needs and tailoring interventions. The precision of IRT models in identifying students' specific abilities and the challenges they face leads to more effective and efficient learning experiences. This study highlights the potential of IRT in improving educational practices and calls for further exploration of its application in diverse educational contexts to optimize learning outcomes.

REFERENCES

Abedalaziz, N., & Leng, C. H. (2018). The relationship between CTT and IRT approaches in Analyzing Item Characteristics. *MOJES: Malaysian Online Journal of Educational Sciences*, *1*(1), 64-70.

Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., & Ali, N. A. (2022).

Assessment and evaluation of different machine learning algorithms for predicting student performance. *Computational intelligence and neuroscience*, *2022*(1), 4151487. https://doi.org/10.1155/2022/4151487

Becker, A., & Nekrasova-Beker, T. (2018). Investigating the effect of different selected-response item formats for reading comprehension. *Educational Assessment*, *23*(4), 296-317. https://doi.org/10.1080/10627197.2018.1517023

Beggrow, E. P., Ha, M., Nehm, R. H., Pearl, D., & Boone, W. J. (2014). Assessing scientific practices using machine-learning methods: How closely do they match clinical interview performance?. *Journal of Science education and Technology*, *23*, 160-182. https://doi.org/10.1007/s10956-013-9461-9

Bustamante, J. C., & Navarro, J. J. (2022). Technological tools for the intervention and computerized dynamic assessment of executive functions. In *Handbook of Research on Neurocognitive Development of Executive Functions and Implications for Intervention* (pp. 310-339). IGI Global Scientific Publishing. https://doi.org/10.4018/978-1-7998-9075-1.ch014

Cardamone, C. N., Abbott, J. E., Rayyan, S., Seaton, D. T., Pawl, A., & Pritchard, D. E. (2012, February). Item response theory analysis of the mechanics baseline test. In *AIP Conference Proceedings* (Vol. 1413, No. 1, pp. 135-138). American Institute of Physics. https://doi.org/10.1063/1.3680012

Carlson, J. E., & von Davier, M. (2017). Item response theory. *Advancing human assessment: The methodological, psychological and policy contributions of ETS*, 133-178. https://doi.org/0.1007/978-3-319-58689-2

Chen, Y., Li, X., Liu, J., & Ying, Z. (2025). Item response theory—A statistical framework for educational and psychological measurement. *Statistical Science*, *40*(2), 167-194. https://doi.org/10.1214/23-STS896

Cook, D. A., & Hatala, R. (2016). Validation of educational assessments: a primer for simulation and beyond. *Advances in simulation*, *1*, 1-12. https://doi.org/10.1186/s41077-016-0033-y

Cook, D. A., Kuper, A., Hatala, R., & Ginsburg, S. (2016). When assessment data are words: validity evidence for qualitative educational assessments. *Academic Medicine*, *91*(10), 1359-1369. https://doi.org/10.1097/ACM.0000000000001175

Csapó, B., & Molnár, G. (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia system. *Frontiers in psychology*, *10*, 1522. https://doi.org/10.3389/fpsyg.2019.01522

Devayanti, D., Suryana, D., & Sunarya, Y. Implementing Rasch Model as an Approach to Test Academic Integrity Instrument's Validity and Reliability. Pedagogia Jurnal Ilmu Pendidikan, 21(1), 25–36. https://doi.org/10.17509/pdgia.v21i1.54133

Diaz, N. V. M., Yoon, S. Y., Trytten, D. A., & Meier, R. (2023). Development and Validation of the Engineering Computational Thinking Diagnostic for Undergraduate Students. *IEEE Access*, *11*, 133099-133114. 10.1109/ACCESS.2023.3335931

Dumont, H., & Ready, D. D. (2023). On the promise of personalized learning for educational equity. *Npj science of learning*, *8*(1), 1-6. https://doi.org/10.1038/s41539-023-00174-x

Eren, B., Gündüz, T., & Tan, Ş. (2023). Comparison of methods used in detection of DIF in cognitive diagnostic models with traditional methods: Applications in TIMSS 2011. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(1), 76-94. https://doi.org/10.21031/epod.1218144

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023). Modeling item-level heterogeneous treatment effects with the explanatory item response model: Leveraging large-scale online assessments to pinpoint the impact of educational interventions. *Journal of Educational and Behavioral Statistics*, *48*(6), 889-913. https://doi.org/10.3102/10769986231171710

Hitt, D. H., & Tucker, P. D. (2016). Systematic review of key leader practices found to influence student achievement: A unified framework. *Review of educational research*, *86*(2), 531-569. https://doi.org/10.3102/00346543156149

Ju, G. F., & Bork, A. (2005, July). The implementation of an adaptive test on the computer. In Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05) (pp. 822-823). IEEE. https://doi.org/10.1109/ICALT.2005.274

Kubsch, M., Czinczel, B., Lossjew, J., Wyrwich, T., Bednorz, D., Bernholt, S., ... & Rummel, N. (2022, August). Toward learning progression analytics—Developing learning environments for the automated analysis of learning using evidence centered design. In *Frontiers in education* (Vol. 7, p. 981910). Frontiers Media SA. https://doi.org/10.3389/feduc.2022.981910

Larrain, M., & Kaiser, G. (2022). Interpretation of students' errors as part of the diagnostic competence of pre-service primary school teachers. *Journal für Mathematik-Didaktik*, *43*(1), 39-66. https://doi.org/10.1007/s13138-022-00198-7

Lee, Y. (2019). Estimating student ability and problem difficulty using item response theory (IRT) and TrueSkill. *Information Discovery and Delivery*, *47*(2), 67-75.

Mallillin, L. L. D. (2022). Teaching and learning intervention in the educational setting: adapting the teacher theory model. *International Journal of Educational Innovation and Research*, *1*(2), 99-121. https://doi.org/10.31949/ijeir.v1i2.2493

Pak, K., Polikoff, M. S., Desimone, L. M., & Saldívar García, E. (2020). The adaptive challenges of curriculum implementation: Insights for educational leaders driving standards-based reform. *Aera Open*, *6*(2), 1 –15. https://doi.org/10.1177/233285842093282

Pea, R. D. (2018). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. In *Scaffolding* (pp. 423-451). Psychology Press.

Pliakos, K., Joo, S. H., Park, J. Y., Cornillie, F., Vens, C., & Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers & Education*, *137*, 91-103. https://doi.org/10.1016/j.compedu.2019.04.009

DOI: https://doi.org/10.17509/pdgia.v22i3.77064
p- ISSN p. ISSN 1693-5276 e- ISSN 2579-7700

Raykov, T., & Marcoulides, G. A. (2016). On the relationship between classical test theory and item response theory: From one to the other and back. *Educational and psychological measurement*, *76*(2), 325-338. https://doi.org/10.1177/0013164415576958

Tetzlaff, L., Schmiedek, F., & Brod, G. (2021). Developing personalized education: A dynamic framework. *Educational Psychology Review*, *33*, 863-882. https://doi.org/10.1007/s10648-020-09570-w

Tian, X., Han, X., Cheng, H. N., Chang, W. C., Liao, C. C., Sun, J., ... & Liu, S. (2017, July). Applying item response theory to analyzing and improving the item quality of an online Chinese reading assessment. In *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)* (pp. 754-759). IEEE. https://doi.org/10.1109/IIAI-AAI.2017.100

Tornabene, R. E., Lavington, E., & Nehm, R. H. (2018). Testing validity inferences for Genetic Drift Inventory scores using Rasch modeling and item order analyses. *Evolution: Education and Outreach*, *11*, 1-16. https://doi.org/10.1186/s12052-018-0082-x

Van der Linden, W. J., & Glas, C. A. (Eds.). (2010). *Elements of adaptive testing* (Vol. 10, pp. 978-0). New York: Springer. https://doi.org/10.1007/978-0-387-85461-8

Willis, L., Badrinarayan, A., & Martinez, M. (2022*). Quality criteria for systems of performance assessment for school, district, and network leaders*. Learning Policy Institute. https://doi.org/10.54300/439.730

DOI: https://doi.org/10.17509/pdgia.v22i3.77064
p- ISSN p. ISSN 1693-5276 e- ISSN 2579-7700