

Innovative Analysis of Mathematics Item Quality Using Item Response Theory for Inclusive and Sustainable Learning

Submitted 11 August 2026, Revised 24 September 2026, Accepted 24 September 2026

Endang Wati^{1*}, Heri Retnawati², Shen Huayu³, Titus Kurra⁴, Muhammad Salman Arrosyid⁵

^{1,5}Department of Educational Research and Evaluation, Graduate School,
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia.

²Department of Mathematics Education, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia.

³Research Institute for International and Comparative Education, Shanghai Normal University, Shanghai, China

⁴Technology and Vocational Education, Faculty of Education, Universitas Stella Maris Sumba, Sumba, Indonesia.
Correspondence Email: *endangwati.2024@student.uny.ac.id

Abstract

High-quality assessment instruments ensure that education measures learning outcomes validly, reliably, and fairly. This study analyzes the relationship between mathematics item quality and student learning outcomes using Item Response Theory (IRT). This data-based approach provides more accurate and comprehensive parameter estimation than classical test theory. The study involved 189 students, and the analysis began with the KMO and Bartlett's Test, which showed sufficient feasibility to proceed to factor exploration. It is clear that the rotation method generated items that fulfilled the requirements of feasibility, making them suitable to be indicators of mathematical competences. Through exploratory factor analysis, it became possible to identify the underlying structure of cognitive constructs of students. This confirms that the developed assessment tools, based on Item Response Theory, facilitate the diagnostics of learning achievements more accurately and help make the process of learning inclusive by providing easily measurable assessments. The results of the study go beyond purely technical aspects since it is important to have well-developed assessment tools to ensure proper measuring of learning outcomes.

Keywords: Assessment, Item Response Theory, Mathematics

INTRODUCTION

The science of mathematics is a field that is very instrumental in the development of logical, analytical, and critical thinking of the students at all educational levels. The study of mathematics is a strategy that has always been incorporated into the national curriculum for years due to its role in developing higher-order thinking and problem-solving skills (Wahono et al., 2026; Septripiyani & Novtiar, 2021). Challenges during mathematics learning are, however, very prevalent, considering how some students fail to grasp the concepts and perform poorly in their learning outcomes (Putri & Rosilowati, 2024). Such a situation points to the challenges that exist in the effective implementation of meaningful learning, and one of those challenges is associated with the assessment process.

Quality learning cannot be separated from good and measurable evaluation activities (Utami et al., 2026; Setiani et al., 2025; Alfarisa & Purnama, 2019). In learning outcome assessment, teachers need evaluation instruments that accurately and fairly describe students' abilities. Two main approaches are classical test theory (CTT) and item response theory (IRT). CTT assesses learning outcomes based on total scores and tends to depend on the tested group,

making item parameters and student abilities relative to the sample (Djemari, 2012; Sudjana, 2016). In contrast, item response theory is a modern approach that explains the probability of correct answers based on students' latent abilities and item characteristics, making it more stable, accurate, and less biased (Hambleton & Swaminathan, 1985; Retnawati, 2014).

The creation of measurement theories including item response theory (IRT) is evidence of transition from score collection to measuring ability using probabilistic models. According to Allen & Yen (1979) and Sari et al., (2022), any phenomenon in educational measurement has to be expressed quantitatively in order to make an objective description of reality. The IRT suggests the use of a mathematical function to correlate the ability of students (latent trait) to the likelihood of providing the right answer and estimation of three parameters – discrimination, difficulty and probability of guessing. This approach provides teachers with better understanding of the characteristics of items and makes decisions data-driven.

The key advantage of item response theory (IRT) over classical test theory is the property of invariance of parameters – item parameters do not depend on the population of respondents and student ability estimates do not depend on a particular set of items (Retnawati, 2014). The item response theory can be used as a method of creation of fair and inclusive educational assessment, when every student has the same chance to be assessed objectively. In addition, item response theory (IRT) supports sustainable learning because accurate measurement data can serve as a basis for teachers to carry out targeted interventions, remediation, or enrichment (Istiyono et al., 2024).

In terms of both theory and practice, quality assessment tools are key components in enhancing student performance. Valid and reliable assessment tools enable teachers to have a clear picture of their students' strengths and weaknesses (Emerentian et al., 2022; Nurjannah et al., 2023). In turn, such knowledge promotes the use of personalized learning strategies. Several previous studies have shown that low-quality test items can bias interpretations of learning outcomes and lead to inappropriate educational decisions (Hardianti, 2021; Rina et al., 2024). Therefore, an analytical approach is needed to provide a comprehensive picture of item quality, based not only on final scores but also on psychometric characteristics.

In this context, item response theory (IRT) offers an innovative approach to analyzing item quality and linking it more comprehensively to student learning outcomes. Through IRT, teachers and researchers can determine how well each item measures the intended mathematical ability and evaluate the instrument's effectiveness in describing student ability profiles. The strengths of item response theory (IRT) lie not only in its technical aspects of parameter

estimation but also in its potential to foster more inclusive and accurate student assessment for a sustainable future of education.

Although Item Response Theory (IRT) has been increasingly applied to evaluate the psychometric characteristics of educational assessments, empirical evidence regarding mathematics test items in junior secondary school contexts remains limited. Previous IRT-based mathematics assessment studies have shown that item-level parameters, particularly discrimination and difficulty, provide detailed information for evaluating the quality and functioning of individual test items (Jumini & Retnawati, 2022). Therefore, an empirical analysis of the mathematics assessment administered at SMP Negeri 1 Barat and SMP II Al Abidin Sukoharjo is needed to identify individual-item characteristics and provide evidence to improve the assessment within the educational context examined.

Operationally, the link between Item Response Theory (IRT) and differentiated instruction can be explained through information derived from item parameters and student ability estimates. The discrimination parameter (a) indicates an item's ability to distinguish between students based on proficiency levels, while the difficulty parameter (b) reflects the ability required to answer an item correctly. This information enables the identification of student ability patterns and the precise determination of learning needs. Thus, measurement results serve as a foundation upon which teachers can implement remediation to students that need reinforcement, proper level of difficulty, or enrichment depending on their performance. Such evidence-based approach emphasizes the role of IRT in ensuring quality and inclusive education as per SDG 4.

Based on the above description, this study aims to analyze the relationship between mathematics test item quality and student learning outcomes through the item response theory (IRT) approach. The results of this analysis are expected to contribute to valid, reliable, and adaptive assessment instruments and to form the basis for creating a smart, fair, and sustainable learning ecosystem aligned with Sustainable Development Goal 4 (SDG 4).

METHOD

Research Design and Participants

This study employed a descriptive quantitative design to examine the psychometric characteristics of a mathematics achievement test using Item Response Theory (IRT). In IRT, the association between the latent trait (ability) and the probability that the examinees get the items right is measured (de Ayala, 2009; Kilpatrick et al., 2001). In this study, the participants were 189 Grade VII students at two junior high schools in the Regency of Sukoharjo, Central Java, Indonesia who were the target population of the Grade VII mathematics test. Nevertheless,

since the study used data only from two junior high schools, the results of this study cannot be generalized to the whole population of Grade VII.

The data were collected using the informal consent and cooperation from the participating junior high schools. The authors did not acquire the formal consent from the school authority, formal written parental consent, formal student assent, and formal ethics committee review. This is a limitation of the study since obtaining the consent is an ethical principle when conducting research involving human participants (Allen & Yen, 1979). It is worth noting that the analytical dataset did not contain any personally identifying information of the participants.

Instrument and Data Collection

The study was conducted on 289 grade seven students from two schools, comprising 128 students (67.7%) from one of the junior high schools in Sukoharjo Regency and 62 students (32.3%) from another junior high school in Sukoharjo Regency, Central Java. This number indicates the total number of students that met the inclusion criteria and were accessible at those respective schools. The present study was not meant to compare the performance of students across the two different schools but used responses of students from both the schools to develop psychometric properties of the mathematics test based on the 2PL IRT model. Hence, the unequal number of participants is an aspect of the sample rather than a reason for comparison between the schools. However, this unequal distribution and sample size should be taken into consideration while evaluating the generalizability and stability of the item parameters (Jumini & Retanawati, 2022).

The tool was made up of 30 items, which were multiple-choice questions with a dichotomous scoring system where 1 was given for a correct response, and 0 was given for wrong answers. The items were derived from Grade VII mathematics competency indicators and were aimed at assessing conceptual and procedural skills, which is evidence for the validity of scores. Prior to the test, the items underwent scrutiny by four experts, two mathematics teachers and two measurement experts. The review examined content relevance, item construction, and language clarity, providing evidence concerning the appropriateness and representativeness of test content (Schmitz & Storey, 2020; Plake & Wise, 2014). The instrument was then administered to 189 participants. No clinical or diagnostic data, including diagnosis of dyslexia, were gathered and used as a criteria for choosing participants. Thus, the results obtained are related to psychometric properties of the test in relation to participating students, but not to students having clinical learning disabilities.

Data Analysis

The process of data analysis entailed two steps. Initially, Kaiser-Meyer-Olkin (KMO) test and Bartlett's Test of Sphericity were used to determine if the correlation matrix was appropriate for factor analysis. While KMO test checks for sampling adequacy, Bartlett's Test checks if the correlation matrix is different from an identity matrix (Kaiser, 1970; Bartlett, 1950). Both KMO test and Bartlett's Test were considered as indicative rather than confirmatory measures of unidimensionality. Second, the test items were tested using the two parameter logistic (2PL) model of Item Response Theory (IRT) (de Ayala, 2009; Kilpatrick et al., 2001):

$$P_i(\theta) = \frac{e^{D_i(\theta-b_i)}}{1+e^{D_i(\theta-b_i)}} \dots\dots\dots (1)$$

Description:

- $P_i(\theta)$: The probability that test-takers with the ability to answer item i will do so
- θ : Test takers' ability level
- b_i : Difficulty index of item i
- E : Natural number whose value is close to 2.718
- D : Scaling factor whose value is close to 1.7

The discrimination parameter indicates how well an item differentiates students by ability, while the difficulty parameter indicates the item's location on the ability scale, helping interpret item performance (de Ayala, 2009; Kilpatrick et al., 2001). We identified items with discrimination parameters below 0.50 for further examination rather than automatically eliminating them. Item decisions were considered together with statistical results and the substantive characteristics of the items, consistent with principles of test development and score interpretation (Plake & Wise, 2014). Pearson's product-moment correlation was also used to examine the relationship between individual item scores and total test scores.

Trustworthiness of the Analytical Process

To improve procedural transparency, the researchers documented the main stages of instrument development, expert review, data collection, coding, screening, and statistical analysis. These records traced the analytical process from the collected responses to the reported results.

RESULTS AND DISCUSSION

At the preliminary step of dimensionality assessment, response matrix was subjected to Kaiser-Meyer-Olkin (KMO) criterion and Bartlett's Test of Sphericity. Obtained KMO index was equal to 0.824, and significant value of χ^2 of Bartlett's Test was 881.562, $p < .001$ (Figure 1). It could be stated that the correlation matrix was suitable for further factor analysis. This finding implied the possibility of factorability, but not unidimensionality (Kaiser, 1974). Further on, exploratory factor analysis (EFA) and scree plot analysis were applied to determine the latent dimensionality of the measuring instrument (Figure 2). Scree plot analysis revealed a single dominant factor and the decrease of eigenvalues after it, thus confirming a dominant latent dimension. Such findings allowed conducting unidimensional 2PL IRT analysis, whereas dimensionality should be interpreted from the overall factor structure and construct definition.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.824
Bartlett's Test of Sphericity	Approx. Chi-Square	881.562
	df	210
	Sig.	.000

Figure 1. Results of KMO and Bartlett's Test

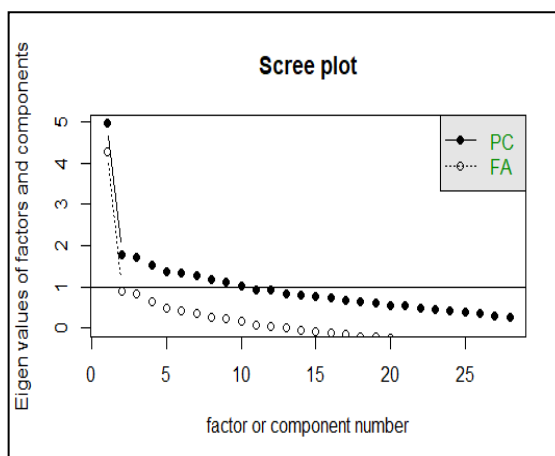


Figure 2. Scree Plot Results of Not Unidimensional Exploratory Factor Analysis (EFA)

The output from the EFA revealed that the vast majority of the items were strongly related to the main factor; but the items 5, 6 and 16 had relatively lower loadings on the main factor. This relatively lower loading implies that these items may contribute less significantly to the main construct compared to the other items. (Hair et al., 2019). Higher factor loadings mean that these items are able to represent the dimensions of the construct effectively. However, you should not rely solely on factor loading as evidence of the need to delete an item. This integrated approach is important because factor loading and IRT discrimination provide different forms of psychometric evidence: factor loading reflects an item's association with the factor structure, whereas the IRT discrimination parameter describes how strongly it

differentiates examinees along the latent ability continuum. Therefore, decisions concerning item revision or removal should consider converging evidence from construct-level and item-level analyses. Factor loading values for each test item based on the results of an exploratory factor analysis (EFA) can be seen in Figure 3.

BUTIR_1	BUTIR_2	BUTIR_3	BUTIR_4	BUTIR_5	BUTIR_6	BUTIR_7	BUTIR_8
BUTIR_9	BUTIR_10						
0.73	0.75	0.78	0.82	0.34	0.56	0.44	0.85
0.78	0.85						
BUTIR_11	BUTIR_12	BUTIR_13	BUTIR_14	BUTIR_15	BUTIR_16	BUTIR_19	BUTIR_20
BUTIR_21	BUTIR_22						
0.59	0.83	0.87	0.85	0.82	0.46	0.70	0.47
0.48	0.53						
BUTIR_23	BUTIR_24	BUTIR_25	BUTIR_26	BUTIR_27	BUTIR_28	BUTIR_29	BUTIR_30
0.54	0.29	0.50	0.81	0.54	0.59	0.82	0.73

Figure 3. Factor Loading Values for Each Test Item based on the Results of an Exploratory Factor Analysis (EFA)

The estimated discrimination parameters ranged between 0.52 to 1.84, suggesting differences in the degree of differentiation of students on the latent mathematics ability continuum by the items. High discrimination parameters lead to steep item characteristic curves around the location of item difficulty, and hence, the probability of a correct response varies with examinee ability. On the other hand, low discrimination parameters provide less differentiation of examinees. This suggests variation in the item-level information that cannot be appropriately represented in a single total-test score. This finding is in line with previous studies conducted on mathematics assessments through 2PL IRT analysis, in which the discrimination parameter at the item level suggested that some of the items needed more attention even as the test functioned effectively (Jumini & Retnawati, 2022).

The estimated difficulty parameters ranged between -1.20 to 1.75 , indicating dispersion of the items across the latent ability continuum. Difficulty parameters with negative values meant that items fell on the lower side of the ability scale, while positive difficulty parameters indicated placement of items towards the higher end of the ability scale. In essence, the test did not cover a narrow range of the latent ability continuum but covered all across the continuum.

As shown by the test information function, the test was found to have the highest measurement information within the approximately $\theta = -1$ to $+1$ range. Therefore, the test is found to be most accurate in the case where the examinees' mathematics ability is within this range. The test information in the IRT is different at each ability level. This means that a test may provide a more accurate measurement of some ability levels than other levels. The high level of information within the range of $\theta = -1$ to $+1$ indicates that this current test is highly informative particularly at the center of the ability continuum, while the accuracy level could be low at the extremes (Trakas, 1997; Jumini & Retnawati, 2022; CASRAI, 2026). Therefore,

the TIF adds to the item parameters through the illustration of the precision of the test and item functioning (Embretson & Reise, 2000; de Ayala, 2009).

The Pearson product-moment correlation analysis yielded a statistically significant positive association, $r = .671$, $p < .001$. The coefficient indicates a substantial positive association within the present sample. However, this finding should be interpreted as association rather than causal evidence. The correlation analysis provides complementary empirical evidence concerning the variables examined. Still, it does not by itself establish that improvements in item quality cause improvements in the accuracy of student ability estimates or learning outcomes. Therefore, interpret the results in conjunction with the item-level IRT findings and the intended interpretation and use of the assessment scores.

The results of this study reinforce the findings that the application of item response theory (IRT) provides significant advantages over classical test theory (CTT) in evaluating item quality and student ability (Hambleton & Swaminathan, 1985; Retnawati, 2014). The fundamental difference between the two lies in parameter independence. In classical test theory (CTT), item scores are interdependent, whereas in item response theory (IRT), item scores are independent, so that the addition or removal of items does not affect the characteristics of other items (Istiyono, 2020; Naga, 1991). These findings indicate that item response theory (IRT) is superior in maintaining measurement stability and providing more accurate ability estimates.

The results of exploratory factor analysis (EFA) show that the Kaiser-Meyer-Olkin (KMO) value of 0.824 and factor loadings mostly above 0.5 reinforce the construct validity of the developed instrument. A high Kaiser-Meyer-Olkin (KMO) value indicates that the correlations between items are quite strong, and the data are suitable for factor analysis. This is in accordance with the criteria proposed by Field (2013) and Kaiser (1974). This condition confirms that the remaining items have good internal consistency and are able to represent the construct of mathematical ability measured stably. The increase in the Kaiser-Meyer-Olkin (KMO) value after the removal of invalid items also indicates that the reduction of items with low contribution has a positive effect on the overall structure of the instrument. This shows that the item selection process is effective in strengthening the dimensions of the construct being measured. Thus, the resulting instrument is more focused, efficient, and has better discriminatory power against variations in student abilities.

The results of discrimination (a) and difficulty (b) estimates present an evenly distributed range, meaning that this testing instrument contains high-quality questions that are able to adequately differentiate the ability levels of students along the whole spectrum of abilities. The even distribution of these parameters means that this testing instrument does not contain either

only easy or difficult questions but also includes different levels of difficulty to ensure a comprehensive measurement of abilities of students. These results correspond to the principle of test information function (TIF), which demonstrates the highest efficiency of measurements in the case of moderate ability levels. Thus, this testing instrument is not only efficient enough as a general tool for measurement but also highly diagnostic. It is extremely important for teachers to use this information in order to develop the optimal strategy of interaction with each student.

From an inclusive education perspective, the application of item response theory (IRT) ensures that the process is fair and unbiased towards differences in students' abilities and backgrounds. The principle of item parameter invariance in IRT ensures that the characteristics of an item do not depend on the test-taker population, so that the instrument can be used for all student groups equally. This means that the level of difficulty and discriminating power of the items remain consistent even when applied to students with different abilities, genders, or social contexts. This condition reflects the application of the principle of fairness in assessment, where every student has the same opportunity to demonstrate their true abilities without being influenced by external factors. Thus, IRT not only functions as a technical approach to measurement but also as an ethical foundation for creating an objective, inclusive, and equitable evaluation system in the context of modern education (Ntumi 2025; Retnawati, 2014).

Reviewed from the perspective of learning sustainability, instruments developed through the item response theory (IRT) approach play an important role in supporting data-driven learning practices. The assessment results obtained from IRT analysis can be used by teachers to adjust learning strategies, whether in the form of remediation, enrichment, or differentiation based on student profiles. Thus, assessment not only functions as a tool for evaluating learning outcomes but also becomes a source of information that strengthens the cycle of continuous and reflective learning. The adaptive use of assessment data helps create an education system that is more responsive to the needs of students. In line with this, the practice of item response theory (IRT)-based assessment supports the mission of Sustainable Development Goal 4 (SDG 4), which is to ensure inclusive and equitable quality education for all learners and promote learning transformation towards a more sustainable and future-oriented direction (Sandoval-Hernández & Carrasco, 2020; Nagy & Adel, 2025).

Theoretically, this finding enriches educational measurement by showing that item response theory not only offers technical advantages in estimating ability parameters and item characteristics. Moreover, item response theory represents a paradigm shift in how researchers and practitioners understand and apply educational assessment. Through a probabilistic model-

based approach, item response theory provides more accurate, adaptive, and bias-free measurements, so that each student is assessed proportionally to their latent ability. This paradigm emphasizes that assessment not only serves as a tool for measuring learning outcomes but also as an instrument for fair and sustainable learning development. Thus, item response theory (IRT) becomes the conceptual foundation for the development of modern educational evaluation that emphasizes validity, fairness, and future orientation.

CONCLUSION

Based on the analysis, the 2PL Item Response Theory model provided empirical estimates of item discrimination and difficulty for the mathematics assessment administered to 189 Grade VII students. Parameters of discrimination varied from 0.52 to 1.84. However, the difficulty parameters varied from -1.20 to 1.75 , reflecting variations in the discriminating power of items and their ability level requirement. The test generated more information about students who had an ability level of approximately $\theta = -1$ to $+1$. Furthermore, the analysis through Pearson correlation revealed a significant positive relationship between item quality and students' performance ($r = 0.671$, $p < 0.001$). This finding implies that the 2PL IRT is capable of generating more information regarding the quality of mathematics test items under review.

SUGGESTIONS

In future studies, the evidence for the quality and fairness of mathematical assessment tools can be further enhanced by combining exploratory factor analysis and confirmatory factor analysis with IRT. The researchers should carefully analyze the items having low factor loadings along with their IRT estimates and then decide about whether to revise them or drop them from the test battery. Differential item functioning analysis is another area that can be explored in future studies, where researchers will try to find out whether there is any difference in item functioning among certain groups of students. It is also recommended that future studies include large and diverse sample size across various schools, regions, and educational environments. Future research can also focus on various polytomous IRT models, including GRM, and combination of IRT with CAT.

ACKNOWLEDGEMENT

The author wishes to thank the Department of Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta from the bottom of his heart for providing him with academic assistance and help in the successful completion of this research. The author is also grateful to the lecturer of the Item Response Theory subject for the guidance, academic inputs and feedback given to the author during the process of developing and analyzing this research.

Availability of data and material

The datasets generated and/or analyzed during the current study are available from the first author upon reasonable request.

Funding

No Funding on this Research

Ethics approval

The study was approved by Ethics Committee of the Universitas Negeri Yogyakarta, Indonesia on 22 October 2026

Patient consent statement

Informed consent was obtained from all individual participants included in the study.

Permission to reproduce material from other sources

No material from other sources was used in this manuscript. Clinical trial registration This study is not a clinical trial and does not require registration.

Declaration of generative AI and AI-assisted technologies in the writing process During the preparation of this work the author used QuillBot in order to improve the language and readability. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Credit authorship contribution statement

Endang Wati, Heri Retnawati, Shen Huayu, Titus Kurra, Muhammad Salman Arrosyid: Writing–review&editing, Writing-originaldraft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formalanalysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no competing financial or non-financial interests in relation to this study.

REFERENCES

- Alfarisa, N., & Purnama, D. (2019). Evaluasi proses pembelajaran matematika di sekolah menengah pertama. *Jurnal Pendidikan Matematika*, 13(1), 34–45.
- Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Brooks/Cole.
- de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. The Guilford Press.
- Djemari, M. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Nuha Medika.
- Emerentian, P., Lestari, W., & Pranoto, Y. K. L. (2022). Development of Character Assessment Instruments for Elementary School Students in Sikka Regency in Mathematics Subjects Materials for Flat Shapes Based on Android Applications. *JERE*, 11(2), 114–127.

- Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). Sage Publications.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis* (8th ed.). Cengage.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Kluwer-Nijhoff Publishing.
- Hardianti, R. (2021). Analisis kualitas butir soal ujian akhir semester pada mata pelajaran matematika. *Jurnal Evaluasi Pendidikan*, 12(1), 55–68.
- Istiyono, E. (2020). *Teori respons butir dan implementasinya dalam asesmen pendidikan*. UNY Press.
- Istiyono, E., Supahar, & Pebriani, I. N. (2024). *Teori Respon Butir Contoh dan Aplikasinya pada Pembelajaran dan Penelitian*. UNY Press.
- Jumini, S., & Retnawati, H. (2022). Estimating Item Parameters and Student Abilities: An IRT 2PL Analysis of Mathematics Examination. *AL-ISHLAH: Jurnal Pendidikan*, 14(1), 385–398. <https://doi.org/10.35445/alishlah.v14i1.926>
- Kaiser, H. F. (1974). An Index of Factorial Simplicity. *Psychometrika*, 39(1), 31–36. <https://doi.org/10.1007/BF02291575>
- Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding It Up: Helping Children Learn Mathematics*. National Academy Press.
- Naga, D. S. (1991). *Teori skor pada pengukuran pendidikan*. Gunadarma.
- Nagy, E. S. A., & Adel, H. M. (2025). Does changing the requirements of measurement types and scales used in survey-based research yield different results? An SDG4-aligned quality management experimental study. *Management & Sustainability: An Arab Review*, 1-25.
- Ntumi, S. (2025). Advanced Multidimensional Item Response Theory Modeling for High-Stakes, Cross-Disciplinary Competency Assessments in Sub-Saharan Africa: A Psychometric Approach to Equity, Adaptivity, and Policy Integration.
- Nurjannah, S., Istiyono, E., Hastuti, W., Iqbal, M., & Kamal, S. (2023). The Application of Aiken's V Method for Evaluating the Content Validity of Instruments that Measure the Implementation of Formative Assessments. *JRERE*, 12(2), 125–133.
- Plake, B. S., & Wise, L. L. (2014). What Is the Role and Importance of the Revised AERA, APA, NCME Standards for Educational and Psychological Testing? *Educational Measurement: Issues and Practice*, 33(4), 4–12. <https://doi.org/10.1111/emip.12045>
- Putri, D. A., & Rosilowati, E. (2024). Faktor-faktor penyebab rendahnya hasil belajar matematika siswa SMP. *Jurnal Pendidikan Dan Pembelajaran*, 14(1), 22–31.
- Retnawati, H. (2014). *Teori Respons Butir dan Penerapannya*. Nuha Medika.

- Rina, D., Syahrir, M., & Arif, M. (2024). Analisis kebutuhan guru terhadap instrumen asesmen berbasis IRT di sekolah menengah pertama. *Jurnal Evaluasi Pendidikan*, 15(1), 77–91.
- Sandoval-Hernández, A., & Carrasco, D. (2020, March). A measurement strategy for SDG thematic indicators 4.7. 4 and 4.7. 5 using international large scale assessments in education. In *Measuring sustainable development knowledge and skills: are there yet?: Methodological progress on SDG Target 4.7*. UNESCO Institute for Statistics.
- Sari, I. J., Pongsophon, P., Vongsangnak, W., Pimthong, P., & Pitipornatapin, S. (2022). The development of molecular genetics concept test for senior high school students using Rasch analysis. *International Journal of Evaluation and Research in Education*, 11(4).
- Schmitz, K., & Storey, V. C. (2020). Empirical test guidelines for content validity: Wash, rinse, and repeat until clean. *Communications of the Association for Information Systems*, 47(1), 64.
- Septripiyani, N., & Novtiar, C. (2021). Analisis kesulitan belajar matematika siswa SMP. *Jurnal Ilmiah Pendidikan Matematika*, 9(2), 101–113.
- Setiani, H., Ruhiat, Y., & Hendrayana, A. (2025). Needs Analysis for Developing a Digital Evaluation System for Micro-skills Achievement in the Merdeka Belajar-Kampus Merdeka (MBKM) Program. *Jurnal Pendidikan Indonesia Gemilang*, 5(2), 94-106
- Sudjana, N. (2016). *Penilaian hasil proses belajar mengajar*. Remaja Rosdakarya.
- Utami, P. N., Supahar, S., & Arrosyid, M. S. (2026). Quality Assurance of Arabic School Examination Development: A Discrepancy Evaluation in Muhammadiyah Elementary Schools. *Jurnal Pendidikan Indonesia Gemilang*, 6(2), 202-221.
- Wahono, B., Purmanna, A., & Bravo, J. A. M. (2026). The Silent Architecture of Science Learning: A Critical Analysis of Visual Representations in Indonesian Integrated STEM Textbooks. *Jurnal Pendidikan Indonesia Gemilang*, 6(1), 39-55.