IJECA (International Journal of Education & Curriculum Application)
Journal homepage: http://journal.ummat.ac.id/index.php/IJECA
ISSN 2614-3380 | Vol. 8, No. 3, December 2025
https://doi.org/10.31764/ijeca.v8i3.36276

Evaluating the Quality of Mid-Semester Mathematics Summative Assessment in Secondary School: A Psychometric Analysis of Test Items

Yoga Tegar Santosa[1*], Dini Wardani Maulida[1], Juli Ferdianto[1], Sri Sutarni[1], Yulia Maftuhah Hidayati[1]
[1]Department of Mathematics Education, Universitas Muhammadiyah Surakarta, Indonesia
✉ Author Corresponding: a418240010@student.ums.ac.id

ABSTRACT

Mid-semester summative assessments play a crucial role in supporting competency-based learning in the Kurikulum Merdeka. However, existing studies and field practices indicate a persistent gap: teachers rarely conduct systematic psychometric evaluations. Addressing this gap, this study aims to (1) analyze the structure and characteristics of a mid-semester mathematics summative assessment and (2) evaluate the quality of its items based on psychometric criteria within the framework of CTT. Using a mixed-methods sequential exploratory design, data were obtained from two mathematics education experts, two mathematics teachers, and a school principal in an Islamic Integrated Secondary School in Sukoharjo Regency. Data sources included interview transcripts, assessment documents, students' response sheets, and expert validation forms. Qualitative data were analyzed through data reduction, display, and conclusion drawing, while quantitative data were examined using Aiken's V and CTT. The findings reveal that the assessment consisted of 40 multiple-choice items and 5 essay questions, covering Number and Algebra elements of Phase D in the Merdeka Curriculum. The items' content validity was moderate, with strengths in language but weaknesses in cognitive level alignment. Empirical results showed some multiple-choice items were invalid, while all essay questions were valid and reliable ($r = 0.88$). Most items were moderately difficult, with a discrimination index from fair to excellent ($0.3 \leq D \leq 0.8$). However, nearly one-third of distractors in the multiple-choice items did not function well. These results highlight the need for improved item construction and teacher capacity-building to ensure assessments that align with the principles of the Kurikulum Merdeka and support high-quality measurement of student competency.

Keywords: Merdeka Curriculum; Psychometric Analysis; Secondary School; Summative Assessment.

Article History:
Received: 06-11-2025
Revised : 25-11-2025
Accepted: 03-12-2025
Online : 15-12-2025

How to Cite (APA style):
Santosa, Y. T., Maulida, D. W., Ferdianto, J., Sutarni, S., & Hidayati, Y. M. (2025). Evaluating the Quality of Mid-Semester Mathematics Summative Assessment in Secondary School: A Psychometric Analysis of Test Items. *IJECA (International Journal of Education and Curriculum Application)*, *8*(3), 431-449. https://doi.org/10.31764/ijeca.v8i3.36276

1. INTRODUCTION

Secondary school holds a strategic position as a transitional period from the concrete nature of mathematics learning in elementary school to the more abstract and conceptual learning in secondary school (Klee & Miller, 2019). However, various studies indicate that many secondary schools students struggle to understand mathematical concepts deeply, leading to low mathematical thinking skills and poor results in national assessments of numeracy (Ahmad et al., 2024; Andriatna et al., 2024; Retnawati et al., 2017; Wahyuni et al., 2024). Therefore, mathematics learning at the secondary schools level requires focused attention to ensure that the learning

process genuinely develops students' conceptual understanding and higher-order thinking skills. Furthermore, in the context of modern-era secondary schools mathematics learning, the focus is no longer solely on students' procedural mastery of mathematical content but also on the quality of the assessments used (Bahena et al., 2024).

Within the Merdeka Curriculum framework, assessment is positioned as an integral part of mathematics learning, including summative assessments which formally measure learning outcomes and monitor student progress (Feldman, 2025; Kemendikbudristek, 2022). In Indonesia, the most common forms of summative assessment are the Mid-Semester Assessment (*Penilaian Tengah Semester* - PTS) and the End-of-Semester Assessment (*Penilaian Akhir Semester* - PAS) (Fitria et al., 2024). Compared to the PAS, the PTS holds a more strategic role as it provides early feedback to identify misconceptions and improve learning strategies before the end of the semester (Charles, 2023; Mumpuni & Ramli, 2018; Sozer et al., 2019; Xuyen, 2023). Consequently, the quality of the PTS instrument is crucial in determining the accuracy of the information teachers receive for making instructional decisions. Therefore, developing a PTS requires high-quality test items supported by scientific testing through a validation process to ensure the instrument is valid, reliable, and accurately measures the intended competencies (Farida & Musyarofah, 2021; Ishaq et al., 2020).

Nevertheless, field observations reveal a contrasting reality. Based on initial observations and interviews with Grade VII mathematics teachers at an Islamic Integrated Secondary School in Sukoharjo Regency, it was found that all mathematics teachers do not yet employ scientific testing, such as validity and reliability tests, in developing summative questions, particularly for the PTS. Teachers generally construct questions independently based on the material taught, without following a systematic instrument development process. Questions are often created merely to fulfill curriculum demands without in-depth review by the school, which should guarantee assessment quality. The interview results also indicated that teachers lack sufficient knowledge and skills in evaluating item quality, such as understanding discrimination index, difficulty level, and the effectiveness of distractors in multiple-choice questions.

To address this problem, a more systematic and data-driven approach is needed to evaluate the quality of test items. Furthermore, Butakor (2022) recommends psychometric analysis to obtain empirical information regarding item characteristics, thereby making the assessment results more valid, reliable, and fair for all students. In this context, Classical Test Theory (CTT) is a widely used framework supported by various studies in education (Anyawale et al., 2022; Nurjanah et al., 2024; Vincent & Shanmugam, 2020).

Assessment is a systematic process of gathering and interpreting information about student learning achievements (Heil & Ifenthaler, 2023). Based on their purpose, assessments are categorized as diagnostic, formative, and summative (Ghimire, 2021). Summative assessment is particularly important as it is used to evaluate learning outcomes after an instructional period and serves as the basis for academic decisions, such as determining final grades, grade promotion, or graduation (Fitria et al., 2024). In the context of Indonesian education, the implementation of summative assessment is regulated by Permendikbudristek No. 21 of 2022, which states that assessment results are used to evaluate the achievement of learning objectives, provide information on learning progress, and determine student competency mastery. The role of summative assessment is not limited to grading but also functions as a tool for accountability and a basis for evaluating learning effectiveness (Griffin et al., 2014). Therefore, the quality of summative assessment instruments must be assured by fulfilling the criteria of validity, reliability, difficulty level, and discrimination index to ensure accurate and fair results.

The Mid-Semester Assessment (PTS) is a form of summative assessment used to measure student learning achievement at the midpoint of the semester (Hadi et al., 2024). According to the Kemendikbudristek (2022) guidelines, PTS results provide feedback on learning progress, help improve the learning process for the remainder of the semester, and serve as a basis for learning outcome reports. Pedagogically, the PTS acts as a checkpoint that helps teachers assess the effectiveness of their teaching strategies and allows students to reflect on their understanding (Sozer et al., 2019). PTS instruments typically consist of learning outcome tests, either in objective or essay form. To yield valid and reliable data, the development of PTS questions must undergo a review and item analysis process covering aspects of content, construction, language, and psychometric characteristics such as validity, reliability, difficulty level, and discrimination index (Nitko & Brookhart, 2011).

Psychometric analysis is the process of evaluating the quality of individual items within a test instrument to ensure it measures an individual's ability validly and reliably (Elgadal & Mariod, 2021). One widely used framework in school-level educational research is CTT (Anyawale et al., 2022; Nurjanah et al., 2024; Vincent & Shanmugam, 2020). Within CTT, analysis involves examining validity, reliability, difficulty level, discrimination index, and the effectiveness of distractors for each item (Hartati & Yogi, 2019; Kenea et al., 2023; Mahphoth et al., 2021). The examination of item validity and reliability is conducted through empirical analysis of student responses to ensure each item consistently and accurately measures the intended construct. Furthermore, difficulty level analysis aims to determine the proportion of students who answer an item correctly, categorizing items as easy, moderate, or difficult. The calculation of the discrimination index determines the extent to which an item can differentiate between high-ability and low-ability students. Additionally, distractor analysis assesses the effectiveness of options in multiple-choice questions in functioning as intended. Thus, applying psychometric analysis allows test designers to identify and revise ineffective items, thereby improving the overall quality of the PTS and ensuring a fairer and more accurate assessment of student learning achievement.

Research on evaluating the quality of summative assessment instruments in mathematics learning has been conducted by several researchers. Koçdar et al. (2016) assessed the quality of multiple-choice questions based on difficulty and discrimination indices categorized according to Bloom's Taxonomy. Manfaat et al. (2021) reviewed the quality of high school mathematics test items through validity, reliability, discrimination index, difficulty level, and distractor effectiveness. Retnawati (2022); Rahmadani & Hidayati (2023) used the Item Response Theory (IRT) approach to evaluate assessment instrument quality. Additionally, Orhani (2024) highlighted the integration of Bloom's Taxonomy in test construction, while Dewi & Prabowo (2022) analyzed the quality of PTS questions at the secondary school level.

However, these prior studies have several limitations. First, most research focuses more on end-of-semester assessments or large-scale tests, leaving mid-semester assessments (PTS) relatively understudied. Second, studies on PTS at the JHS level, such as that by Dewi & Prabowo (2022), did not include empirical validity testing, thus failing to apply the CTT approach comprehensively. Third, no study has simultaneously combined content validity and empirical validity in analyzing the quality of mathematics PTS instruments at the secondary school level.

Based on these gaps, this study aims to address them by applying a comprehensive CTT-based psychometric analysis, encompassing content validity, empirical validity, reliability, difficulty level, discrimination index, and distractor effectiveness. Scientifically, this research contributes by providing a more complete empirical picture of the quality of mathematics PTS instruments in

secondary school and offering critical evidence to support the development of more valid, reliable, and effective summative instruments for use in the learning context.

Based on the above elaboration, this study has two main objectives: (1) to analyze the form and characteristics of the PTS in mathematics at the secondary school level, and (2) to analyze the quality of PTS test items based on psychometric characteristics, including content validity, empirical validity, reliability, difficulty level, discrimination index, and distractor effectiveness, to obtain a comprehensive overview of the assessment instrument's quality.

2. METHODS

2.1 Research Design

This study employed a mixed-methods approach with a sequential exploratory design (Creswell & Clark, 2017). This design was selected as the research was conducted in two complementary phases. The first phase was qualitative, aiming to identify and describe the forms of the mid-semester summative assessment (PTS) through interviews with teachers and the school principal, as well as a documentation study of the assessment instruments. The second phase was quantitative, aiming to analyze the psychometric quality of the test items using student answer sheets. The results of the qualitative analysis were used as a basis for determining the focus of the quantitative analysis, thereby providing a comprehensive overview of the assessment's characteristics and quality.

2.2 Participants

This study involved two mathematics education experts, two mathematics teachers, and the school principal at Daarul Hidayah Islamic Integrated Secondary School in Sukoharjo. The two experts assessed the content validity of the PTS instrument, while the teachers and principal provided information regarding the test development process, the implemented curriculum, and assessment practices. In the quantitative phase, data were obtained from 36 Grade VII students who took the PTS. A saturation sampling technique was used, as the entire population of students who took the test was analyzed.

2.3 Instruments

The instruments used in this study consisted of semi-structured interview guidelines, a documentation format, and several assessment-related documents obtained through a documentation study. These documents included the mid-semester summative assessment test and the students' response sheets, which served as the primary empirical data for the psychometric analysis, even though the test itself was not developed by the researchers. In addition, an expert validation sheet was employed to assess the content validity of the test items, focusing on alignment with Learning Outcomes, clarity of indicators, appropriateness of cognitive level, and clarity of language and context. The interview guidelines and documentation format were validated by two mathematics education experts, with content validity measured via Aiken's V (Aiken, 1980). The interview guidelines scored 0.83 (high category), and the documentation format scored 0.76 (moderate category) (Aiken, 1985). Minor revisions based on expert feedback involved rewording items in the interview guidelines and adding indicators to the documentation format.

2.4 Data Collection

Data collection in this study was carried out through interviews, documentation studies, and expert validation. The data collection process consisted of two main phases. The first phase was the qualitative phase. In this phase, the researchers conducted interviews with mathematics teachers and the school principal and performed a documentation study by reviewing the mid-semester summative assessment scripts and the test blueprints. This stage aimed to gather information on the types and forms of assessments used, the implemented curriculum, the mathematics topics tested, and the assessment development process. Additionally, the researchers collected students' answer sheets as empirical data for the psychometric analysis of the test items.

The second phase was the quantitative phase. In this phase, the researchers submitted the validation sheet to two mathematics education experts to obtain their assessment of the instrument's content validity. This validation process was conducted to ensure the alignment of the test items with the learning outcomes, indicators, cognitive level, as well as language and contextual aspects. Subsequently, the researchers performed a psychometric analysis of the test items based on the data from the students' answer sheets.

2.5 Validity of Data

In this study, qualitative data were obtained through interviews with mathematics teachers and the school principal, validator comments, and a documentation study of the summative assessment instruments. To ensure data trustworthiness, this study employed source triangulation and methodological triangulation techniques (Miles et al., 2014). Source triangulation was conducted by comparing and verifying information obtained from interviews with mathematics teachers against the results of interviews with the school principal. Meanwhile, methodological triangulation was performed by checking the consistency of the interview findings through document analysis, such as the summative assessment instruments.

2.6 Data Analysis

Data analysis in this study was conducted in two stages: qualitative analysis and quantitative analysis. In the first stage, qualitative analysis was performed on the data from interviews and documentation studies through the process of data reduction, data display, and conclusion drawing (Miles et al., 2014). Data from interviews with mathematics teachers and the school principal, as well as summative assessment documents, were reduced to extract relevant information regarding the form and characteristics of the assessment used. Subsequently, the data were presented narratively and in tables to identify patterns and the assessment's alignment with the curriculum, and then conclusions were drawn to determine the form of the assessment, which served as the basis for the quantitative analysis in the next stage.

In the second stage, quantitative analysis was performed on the data from expert assessments and student answer sheets. The validation results from the two experts were analyzed using Aiken's V (Content Validity Index) to determine the content validity level of each item (Aiken, 1980). Furthermore, empirical data from the students' answer sheets were analyzed using CTT procedures to obtain information on empirical validity, reliability (KR-20 or Cronbach's Alpha), difficulty index, discrimination index, and distractor effectiveness (Allen & Yen, 2001; Ebel & Frisbie, 1991; Nitko & Brookhart, 2011). All operational formulas and interpretation criteria are presented in the Appendix.

3. RESULT AND DISCUSSION

3.1 Type of the Mid-Semester Summative Mathematics Assessment

Document analysis and interviews revealed that the summative assessment for the Grade VII Mid-Semester Assessment consisted of 40 multiple-choice items and 5 essay questions. These were developed with reference to the Merdeka Curriculum for Phase D, specifically the Number and Algebra element (see Table 1). However, the selection of the format and number of items was not based on a needs analysis or pedagogical considerations, but rather followed the pattern used in previous years. This was stated by a teacher: "*We use 40 multiple-choice and 5 essay questions because that's how it has always been. So, every year we just follow the same format as before*." (Teacher A, interview, 2025). This indicates that the principle of diverse assessment formats, as recommended by the Merdeka Curriculum, has not been fully implemented. Instead of selecting assessment formats that reflect the targeted competencies, teachers maintain a traditional format without review. This finding aligns with reports by Halimi & Seridi-Bouchelaghem (2021); Anderson-Levitt (2025), which show that school assessments are often designed based on prior habits rather than competency-based design.

Table 1. Elements and Learning Outcomes Used for Constructing the Assessment

Element	Item Numbers	Item Type	Learning Outcomes
Numbers	1-20	Multiple Choice	Students are able to read, write, and compare integers, rational and irrational numbers, decimal numbers, powers and roots, and numbers in scientific notation. They can apply arithmetic operations to real numbers and provide reasonable estimations or approximations in solving problems (including those related to financial literacy).
	1	Essay	
Algebra	21-40	Multiple Choice	Students are able to recognize, predict, and generalize numerical patterns in the arrangement of objects and numbers. They can express real-life situations in algebraic forms. Students are also able to use algebraic properties (commutative, associative, and distributive) to produce equivalent algebraic expressions.

Furthermore, the cognitive domains addressed in the assessment encompassed C1 (remembering), C2 (understanding), C3 (applying), and C4 (analyzing), with the distribution shown in Figure 2. The proportion of questions for each cognitive domain was 13% (6 items) for C1, 31% (14 items) for C2, 54% (24 items) for C3, and 2% (1 item) for C4. This composition signifies that the assessment predominantly measures low to mid-level cognitive abilities and does not provide sufficient scope for assessing higher-order thinking skills. In contrast, the Merdeka Curriculum encourages assessments that measure reasoning, problem-solving, and analytical abilities as part of the Phase D outcomes (Regina, 2024; Wati et al., 2023). Consistent with this, Hadzhikoleva et al. (2025) also emphasize that summative assessments should ideally cover various cognitive levels to comprehensively represent student competencies. However, the effort to achieve this ideal cognitive composition is not supported by the teachers' capacity to evaluate item quality. The results indicate that the Grade VII mathematics teachers have never performed item quality analysis using psychometric principles, such as validity, reliability, difficulty index, discrimination index, or distractor effectiveness. A statement from one teacher reinforces this finding: "*We have never analyzed our test items in that much detail. We also don't*

really understand those kinds of analyses. Usually, if the test has been used and students' results are not too low, we consider the items to be good enough." (Teacher B, interview, 2025). This condition demonstrates that the assessment development process is not yet based on systematic evaluation; thus, the instrument's quality cannot be empirically assured. This finding is consistent with the research by Kissi et al. (2023), which found that many teachers lack sufficient competence in performing psychometric item analysis, as shown in Figure 1.

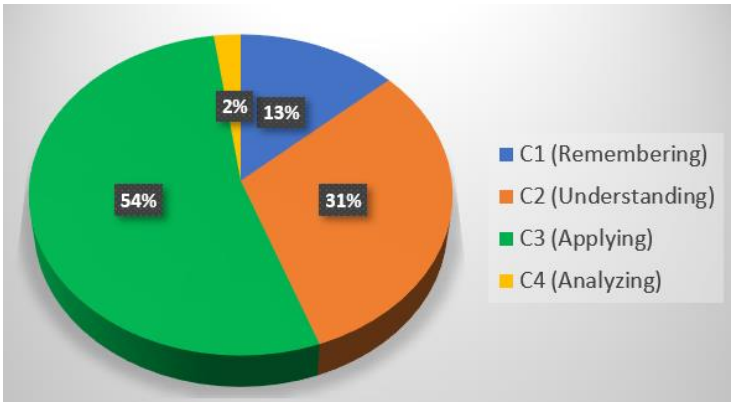

Figure 1. Percentage Distribution of Test Items Across Cognitive Domains

Overall, a summary of the findings for the first objective is presented in Table 2. These findings show that the form, composition, and development practices of the assessments used are not fully aligned with the demands of the Merdeka Curriculum, which emphasizes authentic, diverse assessment oriented towards the development of student competencies. This result strengthens the findings of Alonzo et al. (2021), which showed that teachers' limited understanding of modern assessment principles leads to summative assessments being dominated by old practices with minimal instrument quality evaluation.

Table 2. Summary of Findings and Their Implications for the Merdeka Curriculum

Aspects	Findings	Implications for the Merdeka Curriculum
Assessment Format	40 multiple-choice items and 5 essay items, without needs analysis	Inconsistent with the principles of authentic assessment and varied formats
Cognitive Level	Dominated by C1–C3	Does not support the Phase D Learning Outcomes, which require higher-order thinking skills
Item Development	Based on habitual practices	Does not meet the principle of development-oriented assessment

3.2 Psychometric Quality of the Test Items

a. Content Validity

The content validity analysis results indicate that the quality of the test items is not yet fully aligned with the competencies stipulated in the Merdeka Curriculum. The CVI values, which ranged between 0.33 and 0.833 (see Table 3), reflect inconsistencies in the alignment with indicators, cognitive level, and content representation. The highest scores were in the linguistic aspect, indicating that the items are relatively clear linguistically. However, the lowest scores were in the aspect of cognitive level appropriateness,

suggesting that many items do not measure the thinking skills required for Phase D, particularly reasoning and problem-solving abilities. This finding is reinforced by the qualitative analysis from the validators, who assessed that most items remain in the LOTS (Lower-Order Thinking Skills) domain and tend to be repetitive, thereby reducing their discriminatory power between items.

Table 3. Results of Content Validity Analysis

Aspect	Assessment	CVI	Category
Alignment with Learning Outcomes	The content of the test items aligns with the intended learning outcomes.	0.5	Moderate
	The items are coherent with the elements and objectives of learning within the corresponding phase.	0.667	Moderate
Clarity of Indicators	The indicators reflect the specific competencies being assessed.	0.33	Low
	The indicators and test items are semantically and structurally aligned.	0.5	Moderate
	The indicators are formulated using measurable operational verbs.	0.667	Moderate
Cognitive Level Appropriateness	The cognitive level (C1–C6) corresponds to the required learning outcomes.	0.33	Low
	The variation of cognitive levels is proportionally distributed among easy–moderate–difficult items	0.33	Low
	The test items measure the expected thinking skills (procedural/conceptual/applicative).	0.33	Low
Language and Context Clarity	The language used in the items is clear, communicative, and appropriate to the cognitive level of secondary school.	0.833	High
	The items use meaningful contexts that are relevant and culturally neutral.	0.833	High
	There is no ambiguity or multiple interpretations in the sentence construction	0.833	High

When viewed from a psychometric theory perspective, this finding shows that misalignment in cognitive domains and repetitive item patterns can hinder the instrument's function in validly measuring the variation in student abilities (Pokropek et al., 2022). This condition is also consistent with the findings of Priyatni & Martutik (2020) and Ukobizaba et al. (2021), who reported that school assessments often focus on basic skills and have not yet optimally measured HOTS (Higher-Order Thinking Skills). Thus, the results of this study confirm the need for assessment planning that is more oriented towards higher-order thinking skills (Masyitoh et al., 2020) and systematic review of item quality based on psychometric principles to ensure that the PTS instrument truly reflects the learning outcomes of the Merdeka Curriculum.

b. Empirical Validity

Empirical validity was evaluated to assess the extent to which each test item is consistent with the overall test construct, calculated through the correlation between the item score and the total score. The empirical analysis results for the multiple-choice items (see Table 4) show that the correlation coefficients ranged from −0.276 to 0.597. Psychometrically, this range indicates that the quality of the test items is still varied and tends to be low, as only 12 items (30%) fell into the moderate validity category, while 35% were in the low category and another 35% were invalid. The absence of items with high validity ($r > 0.60$) reflects that most questions are not yet able to consistently represent the measured construct, a condition which, according to Nitko & Brookhart (2019), can weaken the function of summative assessment in accurately depicting learning achievement.

Table 4. Results of Empirical Validity Analysis for Multiple-Choice Items

Item Numbers	Correlation Coefficient Value	Validity Category	Percentage
2, 10, 11, 15, 21, 29, 31, 32, 34, 36, 38, 39	$0.408 \leq r \leq 0.597$	Moderate	30%
4, 6, 7, 12, 13, 14, 16, 18, 19, 22, 23, 24, 27, 28	$0.24 \leq r \leq 0.359$	Low	35%
1, 3, 5, 8, 9, 17, 20, 25, 26, 30, 33, 35, 37, 40	$-0.276 \leq r \leq 0.177$	Invalid	35%

Conversely, the results of the empirical validity analysis for the essay items (see Table 5) show significantly higher correlation coefficients, ranging from 0.679 to 0.820. These values fall into the high to very high categories, indicating that all essay questions have strong and consistent measurement power for the intended construct. This finding aligns with Xiromeriti & Newton (2024), who explained that essay questions are better able to capture higher-order thinking skills such as reasoning, argumentation, and problem-solving which often cannot be adequately evaluated through multiple-choice questions. Therefore, the stark contrast between the validity of multiple-choice and essay items in this assessment confirms an imbalance in instrument quality, particularly in the items' ability to accurately represent the construct. Overall, these empirical validity results demonstrate that the instrument's ability to accurately measure learning outcomes is still limited, especially for multiple-choice items. This condition has direct implications for the quality of summative assessment in schools: an instrument with many invalid items risks generating misleading information about student abilities and undermines the assessment's role as a basis for academic decision-making (Raykov & Zhang, 2025; Roach, 2025).

Table 5. Results of Empirical Validity Analysis for Essay Items

Item Numbers	Correlation Coefficient Value	Validity Category
1	0.739	High
2	0.679	High
3	0.820	Very High
4	0.813	Very High
5	0.775	High

c. Reliability

Reliability testing was conducted to assess the internal consistency of the instrument, which is a crucial indicator for ensuring that summative assessments yield stable and trustworthy information. Based on the analysis results in Table 6, the reliability of the multiple-choice questions was 0.58, while the reliability of the essay questions reached 0.88. This finding indicates a significant difference in consistency between the two question formats. The reliability value of 0.88 for the essay questions reflects high internal consistency, consistent with the view of Nitko & Brookhart (2019) that question formats demanding conceptual understanding and reasoning tend to have better reliability because they provide broader scope for authentic variations in student performance.

In contrast, the reliability value of 0.58 for the multiple-choice questions indicates that this instrument lacks adequate consistency. This suggests that some items are not functioning optimally in measuring the same construct, a condition also reflected in the empirical validity results which showed many items with low correlations. Psychometrically, low reliability indicates potential issues with item structure, misaligned cognitive levels, or content repetition (Malapane & Ndlovu, 2024). Thus, this result reinforces the opinion of Elgadal & Mariod (2021) that the composition and quality of multiple-choice items are not fully aligned with the principles of valid, reliable, and competency-oriented assessment as demanded by the Merdeka Curriculum. In this regard, Zainina et al. (2025) emphasize that teachers need to enhance their capacity for item review and instrument development, especially for multiple-choice formats, so that the PTS can truly yield accurate learning information useful for decision-making.

Table 6. Results of Reliability Analysis for Multiple-Choice and Essay Items

Item Type	Reliability Coefficient	Category
Multiple-Chioce	0.58	Moderate
Essay	0.88	High

d. Item Difficulty Index

The difficulty index analysis was conducted to assess the extent to which the test items proportionally measure student ability. The results in Table 7 show that the multiple-choice questions have a relatively balanced difficulty distribution, with 20% of items classified as difficult, 62.5% as moderate, and 17.5% as easy. This composition is fundamentally in line with psychometric theory recommendations, which suggest a dominance of moderately difficult items as they provide the most optimal measurement information and differentiate student ability more accurately (Popham, 2017). However, the proportion of easy items approaching 20% indicates the need to review some items, particularly to ensure the difficulty level aligns with the competency demands of the Merdeka Curriculum (Marsevani, 2022; Shankar et al., 2024).

Table 7. Difficulty Level of Multiple-Choice Items

Item Numbers	Category	Percentage
1, 5, 9, 12, 17, 26, 37, 40	Difficult	20%
2, 3, 4, 6, 7, 8, 10, 11, 18, 19, 20, 21, 23, 24, 27, 28, 29, 30, 31, 32, 34, 35, 36, 38, 39	Moderate	62.5%
13, 14, 15, 16, 22, 25, 33	Easy	17.5%

For the essay questions, the results in Table 8 show that four items were classified as moderate and one item as easy. The dominance of moderately difficult items reflects a good characteristic, given that essay formats are designed to measure reasoning and conceptual understanding more deeply. Nevertheless, Ginting et al. (2021) suggest that the presence of easy-category items necessitates an increase in complexity for some items to better align with the Phase D learning outcomes of the Merdeka Curriculum, which emphasizes higher-order thinking skills. Overall, these findings indicate that the instrument partially meets the characteristics of an ideal difficulty level but still requires refinement, particularly in composing a more varied difficulty range, to make the assessment more effective in comprehensively measuring the spectrum of student abilities.

Table 8. Difficulty Level of Essay Items

Item Numbers	Category
1	Moderate
2	Moderate
3	Moderate
4	Moderate
5	Easy

e. Item Discrimination Index

The discrimination index test was performed to identify how well each test item differentiates between high-ability and low-ability students. Table 9 shows that the discrimination power of the multiple-choice items varied from poor to excellent. A total of 7 items (17.5%) were in the poor category, 9 items (22.5%) were weak, 12 items (30%) were fair, 11 items (27.5%) were good, and only 1 item (2.5%) reached the excellent category. This finding indicates that while some items have sufficient to good discriminatory power, there are still items that do not function optimally. Items with negative or very low discrimination show an aberrant response pattern, where high-ability students answer incorrectly more often than low-ability students. According to psychometric theory, this condition indicates problems with item construction or ineffective distractors; thus, items in the poor category need to be eliminated, while weak items require revision to improve their discriminatory function (Bhat & Prasad, 2021; Odukoya & Omonijo, 2024).

Table 9. Discrimination Index of Multiple-Choice Items

Item Numbers	Discrimination Index	Category	Percentage
1, 3, 8, 9, 20, 37, 40	$-0.4 \leq D \leq 0$	Poor	17.5%
4, 5, 6, 13, 17, 26, 30, 33, 35	$0.1 \leq D \leq 0.2$	Weak	22.5%
7, 12, 14, 16, 18, 19, 22, 23, 24, 25, 28, 31	$0.3 \leq D \leq 0.40$	Fair	30%
10, 11, 15, 21, 27, 29, 32, 34, 36, 38, 39	$0.5 \leq D \leq 0.7$	Good	27.5%
2	0.8	Excellent	2.5%

Furthermore, Table 10 shows that all essay items had good discrimination power, with four items (numbers 1–4) in the good category (45%–52%) and one item (number 5) in the fair category (37%). This result confirms that the essay question format is more

consistently able to differentiate variations in student ability compared to multiple-choice questions, aligning with the findings of Rush et al. (2016) that essay questions generally have higher discrimination power as they demand more diverse problem-solving strategies and reasoning.

This finding is also consistent with the research of Shakurnia et al. (2022), which reported that low discrimination in multiple-choice questions often arises from non-functional distractors or item construction that does not consider the variation in student ability. The researcher's interviews with mathematics teachers support this finding, indicating that the distractors in the PTS were not designed based on robust conceptual analysis, making some answer options unappealing to low-ability students and unchallenging for high-ability students. Therefore, strengthening teachers' capacity to design effective distractors and review the discriminatory function of items should be a priority in improving the quality of summative assessment.

Table 10. Discrimination Index of Essay Items

Item Numbers	Discrimination Index	Category
1	0.52	Good
2	0.51	Good
3	0.45	Good
4	0.51	Good
5	0.37	Fair

f. Distractor Analysis

The distractor analysis results in Table 11 show that 15 out of 40 multiple-choice items (37.5%) had sub-optimally functioning distractors. Distractors not chosen by any students, as seen in items 3, 7, 12, 14, 28, and 36, indicate that those answer alternatives are either too obviously incorrect or not attractive enough to lure responses from low-ability students. This condition is consistent with the findings of Awalurahman & Budi (2024), who stated that non-functional distractors lead to a disproportionate distribution of answers and negatively impact the item's discrimination power and validity.

Table 11. Analysis of Non-Functional Distractors in Multiple-Choice Items

Item Numbers	Options with Non-Functional Distractors	Item Numbers	Options with Non-Functional Distractors
3	b, c	14	a, c, d
4	a, c	25	c, d
7	a, c	28	d
8	c, d	30	b
9	c	36	b
11	b	39	b
12	a	40	b, d
13	b		

This finding also reinforces the previous psychometric analysis, where items with ineffective distractors tended to show low validity and discrimination values. This aligns with Rezigalla et al. (2024); Terao & Ishii (2020), who argue that distractors that fail to attract low-ability students reduce the item's capacity to differentiate among test-takers'

ability levels. Consequently, more than one-third of the items in this assessment require revision, particularly in the wording of distractors, the plausibility of answer choices, and contextual appropriateness. Such improvements are crucial given the Merdeka Curriculum's emphasis on fair, informative assessment that accurately represents student ability.

Overall, the analysis results indicate that although some components of the instrument (particularly the essay questions) are of good quality, several fundamental weaknesses persist in the multiple-choice items, especially concerning the appropriateness of cognitive domains, empirical validity, discriminatory power, and distractor effectiveness. This finding aligns with the research by Stankous (2016), which showed that mathematical essay questions are more effective in assessing students' conceptual understanding and problem-solving abilities compared to multiple-choice questions.

To provide a more comprehensive overview of the analyzed summative assessment instrument's quality, all psychometric findings from content validity, empirical validity, reliability, difficulty index, discrimination index, to distractor effectiveness are summarized in Table 12. This summary facilitates the observation of consistent patterns across parameters and helps identify the aspects most in need of improvement. Furthermore, presenting this summary helps solidify the relationship between each psychometric indicator and its implications for the overall assessment quality, in accordance with CTT principles and the demands of the Merdeka Curriculum.

Table 12. Summary of Psychometric Evaluation Results and Their Implications

Psychometric Indicator	Main Findings	Implications for Test Quality	Implications for Merdeka Curriculum (MC)
Content Validity (CVI)	1. CVI ranged from 0.33–0.833. 2. Inconsistency in cognitive level alignment. 3. Items dominated by LOTS. 4. Repetitive content.	Indicates weak alignment between items and intended constructs, reduced representativeness, risk of low construct coverage.	Misalignment with MC's emphasis on higher-order reasoning, authentic assessment, and competency-based learning. Highlights need for systematic blueprinting and HOTS-oriented item design.
Empirical Validity	1. MCQs: r = –0.276 to 0.597 (mostly low/invalid) 2. Essays: r = 0.679–0.820 (high–very high).	Many MCQs fail to measure intended construct, inconsistent item–test correlation, essays provide stronger construct validity.	MC requires accurate measurement of competencies, invalid items risk misleading interpretations of student learning. Emphasizes need to strengthen MCQ development practices.

Psychometric Indicator	Main Findings	Implications for Test Quality	Implications for Merdeka Curriculum (MC)
Reliability	1. MCQs α = 0.58 (low) 2. Essays α = 0.88 (high).	Low reliability compromises score stability and test dependability, inconsistent MCQ structure.	MC demands trustworthy assessment evidence for decision-making, reliability weaknesses undermine fairness and accountability. Supports need for teacher capacity-building in item development.
Item Difficulty	1. MCQs: 20% difficult, 62.5% medium, 17.5% easy 2. Essays: mostly medium, one easy.	Overall distribution acceptable but easy items require review, risk of insufficient challenge for competent students.	MC encourages meaningful cognitive challenge aligned with learning progressions, easy items may not reflect required depth of competence.
Discrimination Index	1. MCQs: wide range (poor to excellent) and 40% poor/weak. 2. Essays: all good–fair.	Poor MCQ discrimination signals structural issues, weak distractors, or misalignment, essays better differentiate student ability.	MC promotes assessment that provides diagnostic insights; low discrimination reduces accuracy in distinguishing student mastery levels.
Distractor Functioning	1. 37.5% MCQs have non-functional distractors 2. Several items with distractors never chosen.	Ineffective distractors reduce validity and discrimination, items may become guessable or overly easy.	Violates MC principles of fair, informative, and high-quality assessment, requires strengthening of assessment literacy among teachers.

4. CONCLUSION

This study aimed to analyze the form of the mid-semester summative mathematics assessment and evaluate the quality of its test items based on psychometric characteristics. The results indicate that the instrument consisting of 40 multiple-choice items and 5 essay questions developed based on the Number and Algebra Element in Phase D of the Merdeka Curriculum. The quality of this instrument varies across the different items. Overall, the essay questions have met the criteria for high validity and reliability. In contrast, the multiple-choice items exhibit weaknesses, particularly in the appropriateness of cognitive domains, empirical validity, discriminatory power, and distractor effectiveness. Although most items were at an ideal, moderate difficulty level, the imbalance between LOTS and HOTS questions, coupled with the low internal consistency, indicates a need for improvement in the test development process. Therefore, targeted teacher training in item development and quality analysis is essential to produce assessments that are more accurate, fair, and aligned with the demands of a curriculum based on higher-order thinking skills. Such training should specifically focus on constructing higher-order thinking items, improving distractor writing, and conducting basic psychometric analyses, including validity, reliability, and item discrimination. Given that the greatest weaknesses were found in the multiple-choice items, these aspects should be prioritized. Schools can implement this by organizing internal workshops, collaborating with MGMP, or partnering

with local education authorities to strengthen assessment literacy among teachers. Furthermore, this study is limited by its sample scope, involving only one school; thus, the findings cannot be broadly generalized. Further research is recommended to include a larger number of schools and to investigate the factors influencing teachers' ability to develop high-quality assessment instruments.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Universitas Muhammadiyah Surakarta for full support for this research. Appreciation is also extended to the principal, teachers, and students of SMP IT in Sukoharjo Regency for their participation and cooperation, which significantly contributed to the completion of this study.

REFERENCES

Ahmad, A., Judijanto, L., Jeranah, Halomoan, J. L. A., & Ichsan, M. (2024). Barriers and Difficulties of Students in the Mathematics Learning Process in Junior High Schools. *Journal of Education Research and Evaluation*, *8*(2), 306–316. https://doi.org/10.23887/jere.v8i2.74056

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, *40*(4), 955–959. https://doi.org/10.1177/001316448004000419

Aiken, L. R. (1985). Three Coefficiennts for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, *45*, 131–142. https://doi.org/10.1177/0013164485451012

Allen, M. J., & Yen, W. M. (2001). *Introduction to Measurement Theory*. Waveland Press.

Alonzo, D., Labad, V., Bejano, J., & Guerra, F. (2021). The Policy-driven Dimensions of Teacher Beliefs about Assessment Assessment. *Australian Journal of Teacher Education*, *46*(3), 36–52. https://doi.org/10.14221/ajte.2021v46n3.3

Anderson-Levitt, K. (2025). The deficit model in PISA assessments of competencies: counter-evidence from anthropology. *Globalisation, Societies and Education*, *23*(4), 942–958. https://doi.org/10.1080/14767724.2023.2223141

Andriatna, R., Sujadi, I., Budiyono, Kurniawati, I., Wulandari, A. N., & Puteri, H. A. (2024). Junior high school students' numeracy in geometry and measurement content: Evidence from the minimum competency assessment result. *Proceeding of the 7th National Conference on Mathematics and Mathematics Education (SENATIK)*. https://doi.org/10.1063/5.0194570

Anyawale, M. A., Chere-Masopha, J., & Morena, M. C. (2022). The Classical Test or Item Response Measurement Theory: The Status of the Framework at the Examination Council of Lesotho. *International Journal of Learning, Teaching and Educational Research*, *21*(8), 384–406. https://doi.org/10.26803/ijlter.21.8.22

Awalurahman, H. W., & Budi, I. (2024). Automatic distractor generation in multiple-choice questions: a systematic literature review. *PeerJ Computer Science*, *10*(2), 1–27. https://doi.org/10.7717/peerj-cs.2441

Bahena, R. D., Kilag, O. K. T., Andrin, G. R., Diano, F. M., & Unabia, R. P. (2024). From Method to Equity : Rethinking Mathematics Assessment Policies in Education. *EXCELLENCIA: International Multi-Disciplinary Journal Of Edcation*, *2*(1), 121–132. https://multijournals.org/index.php/excellencia-imje/article/view/281

Bhat, S. K., & Prasad, K. H. (2021). Item analysis and optimizing multiple-choice questions for a viable question bank in ophthalmology. *Indian Journal of Ophthalmology*, *69*(2), 343–346. https://doi.org/10.4103/ijo.IJO_1610_20

Butakor, P. K. (2022). Using Classical Test and Item Response Theories to Evaluate Psychometric Quality of Teacher-Made Test in Ghana. *European Scientific Journal*, *18*(1), 139–168. https://doi.org/10.19044/esj.2022.v18n1p139

Charles, K. J. (2023). Hyflex Instruction: Using Results from Mid-Semester Evaluations for Improvement. *International Journal of Science and Research (IJSR)*, *12*(9), 325–336. https://doi.org/10.21275/SR23901210226

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and Conducting Mixed Methods Research* (3rd ed.). SAGE Publications.

Dewi, W. O., & Prabowo, A. (2022). Item Analysis of the Mid-Semester Assessment for Grade VIII A Mathematics in the 2018/2019 Academic Year at SMP Negeri 3 Mlati. *AdMathEduSt: Jurnal Ilmiah Mahasiswa Pendidikan Matematika*, *9*(2), 76–83. https://doi.org/10.12928/admathedust.v9i2.25347

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of Educational Measurement* (5th ed.). Englewood Cliffs, N.J.

Elgadal, A. H., & Mariod, A. A. (2021). Item Analysis of Multiple-choice Questions (MCQs): Assessment Tool For Quality Assurance Measures. *Sudan Journal of Medical Sciences*, *16*(3), 334–346. https://doi.org/10.18502/sjms.v16i3.9695

Farida, F., & Musyarofah, A. (2021). Validity and Reliability in Item Analysis. *Al-Mu'arrib: Jurnal Pendidikan Bahasa Arab*, *I*(1), 34–44. https://doi.org/10.32923/al-muarrib.v1i1.2100

Feldman, L. I. (2025). The Role of Assessment in Improving Education and Promoting Educational Equity. *Education Sciences*, *15*(2), 1–11. https://doi.org/10.3390/educsci15020224

Fitria, N. N., Mufidah, L. L. N., & Setiawati, P. (2024). Summative Assessment of Islamic Education Subject in Merdeka Curriculum. *Journal of Educational Research and Practice*, *2*(3), 328–338. https://doi.org/10.70376/jerp.v2i3.157

Ghimire, L. (2021). Assessment of the policy. In *Multilingualism in Education in Nepal* (pp. 128–150). Routledge India. https://doi.org/10.4324/9781003159964-7

Ginting, P., Hasnah, Y., Hasibuan, S. H., & Batubara, I. H. (2021). Evaluating Cognitive Level of Final Semester Examination Questions Based on Bloom's Revised Taxonomy. *AL-ISHLAH: Jurnal Pendidikan*, *13*(1), 186–195. https://doi.org/10.35445/alishlah.v13i1.385

Griffin, P., Care, E., Francis, M., & Scoular, C. (2014). The Role of Assessment in Improving Learning in a Context of High Accountability. In *Designing Assessment for Quality Learning* (pp. 73–87). Springer. https://doi.org/10.1007/978-94-007-5902-2_5

Hadi, A. F. M. Q. Al, Listari, D. A., Meilawati, A., & Inayati, N. L. (2024). Implementation of Summative Evaluation in Islamic Education Learning at SMPN 1 Surakarta. *TSAQOFAH*, *4*(1), 769–778. https://doi.org/10.58578/tsaqofah.v4i1.2570

Hadzhikoleva, S., Hadzhikolev, E., Gaftandzhieva, S., & Pashev, G. (2025). A conceptual framework for multi-component summative assessment in an e-learning management system. *Frontiers in Education*, *10*(1), 1–12. https://doi.org/10.3389/feduc.2025.1656092

Halimi, K., & Seridi-Bouchelaghem, H. (2021). Students' competencies discovery and assessment using learning analytics and semantic web. *Australasian Journal of Educational Technology*, *37*(5), 77–97. https://doi.org/10.14742/ajet.7116

Hartati, N., & Yogi, H. P. S. (2019). Item Analysis for a Better Quality Test. *English Language in Focus*, *2*(1), 57–70. https://doi.org/10.24853/elif.2.1.59-70

Heil, J., & Ifenthaler, D. (2023). Online Assessment in Higher Education: A Systematic Review. *Online Learning*, *27*(1), 187–218. https://doi.org/10.24059/olj.v27i1.3398

Ishaq, K., Majid, A., Rana, K., Azan, N., & Zin, M. (2020). Exploring Summative Assessment and Effects : Primary to Higher Education. *Bulletin Of Education and Research*, *42*(3), 23–50. https://eric.ed.gov/?id=EJ1291061

Kemendikbudristek. (2022). *Learning and Assessment in Early Childhood, Primary, and Secondary Education (Pembelajaran dan Asesmen Pendidikan Anak Usia Dini, Pendidikan Dasar, dan Menegah)*.

Kenea, T. G., Mikire, F., & Negawo, Z. (2023). The Psychometric Properties and Performances of Teacher-Made Tests in Measuring Students ' Academic Performance in Ethiopian Public Universities : Baseline Survey Study. *Research Square*, *1*(4), 1–23. https://doi.org/10.21203/rs.3.rs-3095433/v1

Kissi, P., Baidoo-Anu, D., Anane, E., & Annan-Brew, R. K. (2023). Teachers' test construction

competencies in examination-oriented educational system: Exploring teachers' multiple-choice test construction competence. *Frontiers in Education*, *8*(1), 1–14. https://doi.org/10.3389/feduc.2023.1154592

Klee, H. L., & Miller, A. D. (2019). Moving Up! Or Down? Mathematics Anxiety in the Transition From Elementary School to Junior High. *The Journal of Early Adolescence*, *39*(9), 1311–1336. https://doi.org/10.1177/0272431618825358

Koçdar, S., Karadag, N., & Sahin, M. D. (2016). Analysis of the Difficulty and Discrimination Indices of Multiple-Choice Questions According to Cognitive Levels in an Open and Distance Learning Context. *The Turkish Online Journal of Education Technology*, *15*(4), 16–24. https://eric.ed.gov/?id=EJ1117619

Mahphoth, M. H., Sulaiman, Z., Koe, W., Kamarudin, N. A., Puspo, & Dirgantari, D. (2021). Psychometric Assessment of Young Visitors at the National Museum Of Malaysia. *Asian Journal of University Education*, *17*(2), 1–13. https://doi.org/10.24191/ajue.v17i2.13396

Malapane, T. A., & Ndlovu, N. K. (2024). Assessing the Reliability of Likert Scale Statements in an E-Commerce Quantitative Study: A Cronbach Alpha Analysis Using SPSS Statistics. *2024 Systems and Information Engineering Design Symposium (SIEDS)*, 90–95. https://doi.org/10.1109/SIEDS61124.2024.10534753

Manfaat, B., Nurazizah, A., & Misri, M. A. (2021). Analysis of mathematics test items quality for high school Analysis of mathematics test items quality for high school. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *25*(1), 108–117. https://doi.org/10.21831/pep.v25i1.39174

Marsevani, M. (2022). Item Analysis Of Multiple-Choice Questions: An Assessment Of Young Learners. *English Review: Journal of English Education*, *10*(2), 401–408. https://doi.org/10.25134/erjee.v10i2.6241

Masyitoh, M., Ahda, Y., Hartanto, I., & Darussyamsu, R. (2020). An Analysis of High Order Thinking Skills Aspects on the Assessment Instruments Environmental Change Topic for the 10th GradeSenior High School Students. *Jurnal Atrium Pendidikan Biologi*, *5*(4), 1–7. https://doi.org/10.24036/apb.v5i4.6945

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook* (3rd ed.). SAGE Publications. https://www.ucg.ac.me/skladiste/blog_609332/objava_105202/fajlovi/Creswell.pdf

Mumpuni, K. E., & Ramli, M. (2018). Students' Understanding and Approvement toward Assessment for Learning. *BIOEDUKASI: Jurnal Pendidikan Biologi*, *11*(1), 55–60. https://jurnal.uns.ac.id/bioedukasi/article/download/19746/pdf

Nitko, A. J., & Brookhart, S. M. (2011). *Educational Assessment of Students* (6th ed.). Pearson/Allyn & Bacon.

Nitko, A. J., & Brookhart, S. M. (2019). *Educational Assessment of Students* (8th ed.). Pearson.

Nurjanah, S., Iqbal, M., Zafrullah, Z., Mahmud, M. N., Seran, D. S. F., Suardi, I. K., & Arriza, L. (2024). Psychometric quality of multiple-choice tests under classical test theory (CTT): AnBuso, Iteman, and R. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *28*(2), 161–172. https://doi.org/10.21831/pep.v28i2.71542

Odukoya, J. A., & Omonijo, D. O. (2024). Discriminatory indices of 'introduction to psychology' multiple choice examination questions. *Edelweiss Applied Science and Technology*, *8*(6), 8833–8847. https://doi.org/10.55214/25768484.v8i6.3880

Orhani, S. (2024). Preparation of Tests from the Subject of Mathematics According to Bloom ' s Taxonomy. *International Journal of Research Publication and Reviews*, *5*(2), 2335–2345. https://doi.org/10.55248/gengpi.5.0224.0542

Pokropek, A., Marks, G. N., & Borgonovi, F. (2022). How much do students' scores in PISA reflect general intelligence and how much do they reflect specific abilities? *Journal of Educational Psychology*, *114*(5), 1121–1135. https://doi.org/10.1037/edu0000687

Popham, W. J. (2017). *Classroom Assessment: What Teachers Need to Know*. Pearson Education.

Priyatni, E. T., & Martutik. (2020). The Development of a Critical–Creative Reading Assessment Based on Problem Solving. *Sage Open*, *10*(2), 1–9. https://doi.org/10.1177/2158244020923350

Rahmadani, N., & Hidayati, K. (2023). Quality of Mathematics Even Semester Final Assessment Test in Class VIII Using R Program. *Jurnal Pendidikan Matematika*, *17*(3), 397–416. https://doi.org/10.22342/jpm.17.3.20627.397-416

Raykov, T., & Zhang, B. (2025). The One-Parameter Logistic Model Can Be True With Zero Probability for a Unidimensional Measuring Instrument: How One Could Go Wrong Removing Items Not Satisfying the Model. *Educational and Psychological Measurement*, *85*(4). https://doi.org/10.1177/00131644251345120

Regina, A. (2024). Assessment Rubric for Historical Thinking Skills in Accordance with the Kurikulum Merdeka. *EDUTEC : Journal of Education And Technology*, *7*(4). https://doi.org/10.29062/edu.v7i4.784

Retnawati, H. (2022). Estimating Item Parameters and Student Abilities : An IRT 2PL Analysis of Mathematics Examination. *Al-Islah: Jurnal Pendidikan*, *14*(1), 385–398. https://doi.org/10.35445/alishlah.v14i1.926

Retnawati, H., Kartowagiran, B., Arlinwibowo, J., & Sulistyaningsih, E. (2017). Why are the Mathematics National Examination Items Difficult and What Is Teachers' Strategy to Overcome It? *International Journal of Instruction*, *10*(3), 257–276. https://doi.org/10.12973/iji.2017.10317a

Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. (2024). Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, *24*(1), 445–451. https://doi.org/10.1186/s12909-024-05433-y

Roach, V. A. (2025). Validity: Conceptualizations for anatomy and health professions educators. *Anatomical Sciences Education*, *18*(8), 751–756. https://doi.org/10.1002/ase.70016

Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, *16*(1), 250–259. https://doi.org/10.1186/s12909-016-0773-3

Shakurnia, A., Ghafourian, M., Khodadadi, A., Ghadiri, A., Amari, A., & Shariffat, M. (2022). Evaluating Functional and Non-Functional Distractors and Their Relationship with Difficulty and Discrimination Indices in Four-Option Multiple-Choice Questions. *Education in Medicine Journal*, *14*(4), 55–62. https://doi.org/10.21315/eimj2022.14.4.5

Shankar, D. R., Singh, D. H. P., Dewan, D. S., & Singh, D. R. (2024). An In-Depth Analysis of Multiple-Choice Question Quality In Community Medicine Examinations: Evaluating Implications For Competency-Based Medical Education At Noida International Institute Of Medical Sciences (NIIMS). *African Journal of Biomedical Research*, *27*(45), 13959–13964. https://doi.org/10.53555/AJBR.v27i4S.7072

Sozer, E. M., Zeybekoglu, Z., & Kaya, M. (2019). Using mid-semester course evaluation as a feedback tool for improving learning and teaching in higher education. *Assessment & Evaluation in Higher Education*, *44*(7), 1003–1016. https://doi.org/10.1080/02602938.2018.1564810

Stankous, N. V. (2016). Constructive Response Vs. Multiple-Choice Tests In Math: American Experience And Discussion (Review). *European Scientific Journal*, *12*(10), 1–9. https://doi.org/10.19044/esj.2016.v12n10p%p

Terao, T., & Ishii, H. (2020). A Comparison of Distractor Selection Among Proficiency Levels in Reading Tests: A Focus on Summarization Processes in Japanese EFL Learners. *Sage Open*, *10*(1), 1–14. https://doi.org/10.1177/2158244020902087

Ukobizaba, F., Nizeyimana, G., & Mukuka, A. (2021). Assessment Strategies for Enhancing Students' Mathematical Problem-solving Skills: A Review of Literature. *Eurasia Journal of Mathematics, Science and Technology Education*, *17*(3), 1–10. https://doi.org/10.29333/ejmste/9728

Vincent, W., & Shanmugam, S. K. S. (2020). The Role of Classical Test Theory to Determine the Quality of Classroom Teaching Test Items. *Pedagogia : Jurnal Pendidikan*, *9*(1), 5–34. https://doi.org/10.21070/pedagogia.v9i1.123

Wahyuni, A., Muhaimin, L. H., Hendriyanto, A., & Tririnika, Y. (2024). Exploring Middle School Students' Challenges in Mathematical Literacy: A Study on AKM Problem-Solving. *AL-ISHLAH: Jurnal Pendidikan*, *16*(3), 3335–3349. https://doi.org/10.35445/alishlah.v16i3.5729

Wati, D. D. E., Dewi, R. K., & Amri, C. (2023). Analysis of student ability formulating learning objectives in natural science phase D kurikulum merdeka. *Jurnal Atrium Pendidikan Biologi*, *8*(1), 15–21. https://doi.org/10.24036/apb.v8i1.14028

Xiromeriti, M., & Newton, P. M. (2024). Solving Not Answering. Validation of Guidance for Writing Higher-Order Multiple-Choice Questions in Medical Science Education. *Medical Science Educator*, *34*(6), 1469–1477. https://doi.org/10.1007/s40670-024-02140-7

Xuyen, P. T. M. (2023). Exploring the Efficacy of Summative Assessment to Promote the Continuous Improvement of Students' English Proficiency. *US-China Education Review B*, *13*(6), 346–357. https://doi.org/10.17265/2161-6248/2023.06.002

Zainina, K. A., Mufiqoh, M. Z., Aprilia, N., & Isnaeni, B. (2025). Rasch Model: Analysis of Biology Question Item in the Indonesia Independent Curriculum. *Jurnal Penelitian Pendidikan IPA*, *10*(12), 10990–10998. https://doi.org/10.29303/jppipa.v10i12.7661