Jurnal Gantang IX (2) (2024): 165-180

JURNAL GANTANG

e-ISSN: 2548-5547
p-ISSN: 2503-0671

http://ojs.umrah.ac.id/index.php/gantang/index

Harmonizing technology and tradition: Analysis of grade VI mathematics midterm exam questions

Eny Cahyaningsih*, Iva Sarifah, Riyadi

State University of Jakarta, DKI Jakarta, 12310, Indonesia
*Corresponding Author : cahyanigsiheny@gmail.com

Submission: December 1st, 2024; Accepted: December 24th, 2024; Published: December 31st, 2024
DOI: https://doi.org/10.31629/jg.v9i2.7467

Abstract

In education, evaluation instruments such as tests are important tools to measure students' abilities objectively. However, unbalanced question quality, such as too easy difficulty levels or ineffective distractors, can reduce the validity and reliability of the test in distinguishing participants' abilities. This study analyzed the quality of the midterm test Mathematics questions for Grade VI at Sekolah Prestasi Global. Analysis using CTT with the help of SPSS and JMetrik allows for a more detailed evaluation of the quality of the questions. The source of research data was the results of 111 students' answers to 25 Midterm Exam Questions, with 15 multiple-choice questions analyzed further, consisting of 10 multiple-choice questions and five true-false questions. The results showed that most of the UTS Mathematics questions for Grade VI were relatively easy (80% with a difficulty level >0.70), so they were less effective in distinguishing participants' abilities. Only three questions were in the ideal difficulty level category (0.31–0.70), while most distractors were ineffective with low or zero discrimination values. Although the overall reliability of the question was good (Cronbach's Alpha 0.763), improvements in test quality are recommended through question revisions, distractor improvements, and a more balanced distribution of difficulty levels. This study concludes that the quality of the Grade VI Mathematics UTS question at Sekolah Prestasi Global needs to be improved through revision of easy questions, improvement of distractors, and a more balanced distribution of difficulty levels to produce a test instrument that is more valid, reliable, and able to evaluate participants' abilities accurately.

Keywords: question quality, classical test theory, validity, reliability, distractor effectiveness.

How to cite: Cahyanigsih, E., Sarifah, I., & Riyadi. Harmonizing technology and tradition: Analysis of grade VI mathematics midterm exam questions. *Jurnal Gantang*, *9*(2), 165 – 180. https://doi.org/10.31629/jg.v9i2.7464

I. Introduction

In education, test instruments play a crucial role in measuring students' abilities and understanding, which requires a high level of validity and reliability so that the results can be trusted (Anshari et al., 2024; Saputri et al., 2023; Zakiyah & Kartika, 2024). Validity ensures that the instrument measures what it is supposed to measure, while reliability ensures the consistency of the measurement results. Several studies have highlighted the importance of validity and reliability in educational test instruments. For example, Saputri et al. (2023), in the analysis of assessment instruments,

emphasized that validity and reliability are the main characteristics that evaluation instruments must meet to ensure the accuracy and consistency of measurement.

Furthermore, Anshari et al. (2024) analyzed the validity and reliability of the final summative test items of the odd semester of Islamic Religious Education (PAI) subjects. They found that the instrument's validity and reliability determine the instrument's quality, and factors such as response, conditions or circumstances of the research location, and the use of non-ideal tools significantly affect the validation process. Finally, Zakiyah and Kartika (2024) tested the content validity of the mathematical representation ability instrument in solving flat shape problems. The study results show that the instrument has a high validity and robust reliability level, which is suitable for measuring students' mathematical representation abilities. One of the main approaches in evaluating test quality is the Classical Test Theory (CTT), which provides a framework for analyzing question characteristics, internal consistency, and measurement error. According to CTT, test scores reflect a combination of true scores and measurement errors, focusing on measuring specific attributes such as ability and knowledge by analyzing parameters such as validity, reliability, difficulty level, and question discrimination. Despite its limitations, such as population dependence and linearity assumptions, this theory remains an important basis in psychometric and educational measurement, as explained by Crocker & Algina (2008) and Anastasi and Urbina (1997).

Previous research has shown that question analysis is important to ensure that assessment instruments accurately reflect students' abilities (Rasmuin & Luddin, 2022). It emphasizes that an in-depth evaluation of exam questions' difficulty level and discriminatory power can improve the quality of questions teachers create. Kaldaras et al. (2024) discussed integrating learning development and artificial intelligence in STEM education to assess knowledge application, highlighting the importance of sophisticated question analysis in accurately measuring students' understanding.

Furthermore, Suprapto et al. (2020) analyzed the quality of an instrument designed to measure students' higher-order thinking skills in physics learning, reinforcing the need for careful question analysis to ensure the validity and reliability of assessment tools.

This research demonstrates that while fundamental analysis provides a solid foundation, a more in-depth and context-specific evaluation of assessment instruments is crucial for accurately capturing and improving student learning outcomes. Prior studies often focus solely on fundamental analysis, underscoring the need for further efforts to develop and optimize question quality based on specific contexts and requirements. Building on previous research, this study integrates Classical Test Theory (CTT) with modern statistical tools such as SPSS and JMetric to comprehensively evaluate test questions. Key contributions of this study include: (i) Advanced distractor analysis, unlike earlier studies that primarily emphasized correct-answer discrimination, this research assesses distractor effectiveness through manual calculations and JMetric-based metrics. It identifies flaws, such as near-zero discrimination values, that are often overlooked in prior work; (ii) Balanced methodology, by comparing manual and automated approaches, the study highlights JMetric's advantages in speed, precision, and visual analysis, enabling more data-driven evaluations; (iii) Question design refinement, the findings reveal that 80% of the questions were too easy, limiting their ability to differentiate student performance. To enhance measurement accuracy, the study recommends a balanced distribution of difficulty levels (30% easy, 40% moderate, 30% difficult); (iv) Practical framework for improvement, the study proposes strategies for replacing invalid items, enhancing distractors, and incorporating evidence-based tools for future assessments. This scalable framework is adaptable to various

subjects and assessment contexts.

The novelty of this research lies in its deeper focus on distractor analysis, its multi-method approach, and its practical framework for creating valid, reliable, and fair assessments. By addressing gaps in distractor evaluation, discrimination power, and difficulty balancing, this study advances test evaluation practices and promotes the development of more accurate and effective measurement tools in education.

II. Research Method

This study employs a quantitative method with a descriptive approach to analyze the quality of Grade VI Mathematics Mid-Term Exam questions. The respondents were Grade VI students at Sekolah Prestasi Global who participated in the mid-semester summative exam. Sekolah Prestasi Global was selected purposively due to its implementation of the Merdeka curriculum, which emphasizes competency-based evaluation and developing critical thinking skills.

The test instrument, prepared by a team of schoolteachers, consisted of 15 exam questions—10 multiple-choice questions and five true-false questions—drawn from a larger pool of 25 exam questions. These questions covered various mathematical concepts, including arithmetic operations, fractions, decimals, geometry, and problem-solving skills. The primary purpose of this exam was to assess students' comprehension, analytical abilities, and reasoning skills based on the material taught during the semester.

The test preparation process was based on a blueprint that classified the questions according to difficulty levels (easy, medium, and difficult) and cognitive aspects (understanding, application, and analysis). The prepared questions were then analyzed using Classical Test Theory (CTT) to evaluate the validity, reliability, difficulty level, discrimination power, and distractor effectiveness. These metrics ensured that the instrument produced accurate and reliable results.

The analysis was conducted using SPSS and JMetrik software, facilitating calculations such as correlation coefficients, Cronbach's Alpha, and answer choice distributions. The five stages of analysis within the CTT framework included:

1. Validity Testing – Using SPSS and the product-moment correlation method to measure the relationship between each question and the total test score.
2. Reliability Testing – Calculate the Cronbach's Alpha coefficient in SPSS to assess internal consistency.
3. Difficulty Level Analysis – Using JMetrik and manual calculations to determine the correct proportion of students answering each question.
4. Discrimination Power Analysis – Evaluating the effectiveness of questions in distinguishing between high- and low-performing students through JMetrik and manual methods.
5. Distractor Effectiveness Analysis – Assessing distractor performance using JMetrik and manual evaluations to ensure optimal functioning of distractors.

The data processing utilized Microsoft Excel, SPSS, and Jmetrik software to generate coefficients for validity, reliability, difficulty levels, discrimination power, and distractor effectiveness. Analysis results were interpreted using theoretical criteria established by Azwar (2019) and Nunnally & Bernstein (1994), ensuring conclusions aligned with empirical standards.

The results presentation included tables and descriptive statistics highlighting key findings—such as the percentage of valid and reliable items, difficulty distributions, and distractor performance. Visual aids emphasized areas needing improvement and supported recommendations for refining the test instrument. Additionally, theoretical frameworks and empirical data interpretations provided practical recommendations for enhancing test design and evaluation processes.

III. Results and Discussion

CTT is an important measurement method for assessing and analyzing test results (Ilhan, 2016). CTT is a very important approach in psychometrics and educational evaluation. This theory is oriented toward the relationship between the scores obtained by individuals in a test and the actual scores, often called "true scores" (LeBeau et al., 2020). True scores reflect an individual's ability without any influence from external factors or measurement errors. For example, if a student takes a math test and gets an 80, but his actual math ability is 85, then 85 is his true score. Measurement errors can be caused by test anxiety, poor physical condition, or even errors in preparing the test itself. Therefore, minimizing these factors is important so that test results accurately reflect students' abilities. CTT is very important in education because it provides a comprehensive framework for understanding how tests can measure students' abilities and knowledge (Haw et al., 2022). With a deeper understanding of CTT, we can better design, implement, and interpret the results of various assessment forms.

The questions analyzed in this study were taken from the Grade VI Mathematics Mid-Term Exam at Sekolah Prestasi Global. A total of 15 questions were selected for evaluation, consisting of 10 multiple-choice questions and five true-false questions, which were extracted from a complete set of 25 exam questions. The content and topics covered in these questions focused on several key mathematical concepts. Arithmetic operations included tasks related to multiplication, division, addition, subtraction, and simplifying results. Fractions tested students' ability to simplify fractions, convert decimals to fractions, and compare ratios. Decimals addressed decimal representation, conversion, and operations involving decimals, while geometry evaluated knowledge about areas and perimeters of geometric shapes, such as rectangles and squares. Additionally, problem-solving questions involved word problems that required calculating prices, making comparisons, and reasoning proportionally.

The 10 multiple-choice questions assessed computation accuracy, concept application, and problem-solving skills. Examples included calculating the cost of multiple items based on unit price, finding the simplest form of ratios and fractions, and determining the remaining balance after a purchase. Meanwhile, the five true-false questions focused on conceptual understanding and reasoning. These included evaluating statements about geometric properties and verifying the accuracy of simplified ratios and decimal conversions. These questions comprehensively assessed students' mathematical abilities, ranging from basic arithmetic operations to more complex problem-solving tasks, as shown in Table 1.

Table 1. Content focus and descriptions of test questions

Question	Content Focus	Description
Q1	Arithmetic Operations	Simplify the result of a multiplication operation.
Q2	Arithmetic Operations	Find the simplest form of a multiplication result.
Q3	Division and Simplification	Simplify the result of a division operation.
Q4	Fractions	Simplify a fraction to its simplest form.
Q5	Decimals to Fractions	Convert a decimal (e.g., 0.75) into its simplest fraction form.
Q6	Number Comparison	Compare decimal numbers and determine their relative sizes.
Q7	Ratios and Proportions	Simplify a given ratio to its simplest form.
Q8	Problem-Solving (Cost Calculation)	Calculate the total cost of purchasing multiple items based on the unit price.

Q9	Ratio and Budget Comparison	Determine the ratio between total budget and expenses.
Q10	Problem-Solving(Change Calculation)	Calculate the cost and amount of change from a real-life shopping problem.
Q11	Geometry (Area Calculation)	Verify the correctness of an area calculation formula for a rectangle.
Q12	Division and Simplification	Check whether the result of a division operation is correctly simplified.
Q13	Decimal Conversion	Evaluate whether the decimal representation of a given fraction is accurate.
Q14	Ratio and Simplification	Determine whether a ratio presented is simplified correctly.
Q15	Proportional Reasoning (Real-life Scenario)	Verify the simplified ratio of colored balls in a box compared to the total number of balls.

Validity

Validity measures the extent to which a question can measure what should be measured according to learning objectives, including knowledge, skills, and attitudes (Sumintono & Widhiarso, 2015). In classical test theory (CTT), validity is defined as the ability of a test to measure the established construct accurately (Azwar, 2019). Validity evaluation is often done by calculating the Pearson correlation between questions and total scores using a formula that measures the relationship between the two variables. A question is considered valid if its correlation coefficient (rxy) is greater than the r table or the significance value is less than 0.05, indicating a significant relationship with the measured construct. Conversely, invalid questions have low or insignificant correlation values, so they need To be revised or deleted to increase the overall validity of the instrument.

Based on the validity test of the Mathematics Mid-Term Exam Questions for Grade VI at Sekolah Prestasi Global using SPSS, the results showed that 14 out of 15 questions had a correlation value with a total score greater than the R Table value (0.1865) and a significance value of 0.000, as shown in Table 2.

Table 2. Correlation values for each item

Item	Correlation with Total Score	R Table	Significance	Valid/Invalid
Q1	0.614	0.1865	0.000	Valid
Q2	0.459	0.1865	0.000	Valid
Q3	0.540	0.1865	0.000	Valid
Q4	0.656	0.1865	0.000	Valid
Q5	0.731	0.1865	0.000	Valid
Q6	0.440	0.1865	0.000	Valid
Q7	0.572	0.1865	0.000	Valid
Q8	0.529	0.1865	0.000	Valid
Q9	0.639	0.1865	0.000	Valid
Q10	0.340	0.1865	0.000	Valid
Q11	0.432	0.1865	0.000	Valid
Q12	0.360	0.1865	0.000	Valid
Q13	0.182	0.1865	0.056	Invalid*
Q14	0.448	0.1865	0.000	Valid
Q15	0.535	0.1865	0.000	Valid

Therefore, these questions were declared valid. However, Question 13 had a correlation value of 0.182 with a significance of 0.056, rendering it invalid. These results indicate that most questions possess good validity, except for Question 13, which requires revision or removal to improve the overall quality of the instrument.

Further analysis confirmed that all questions, except Question 13, met the validity criteria with an r count > r table (0.1865) or a significance value < 0.05. The low correlation observed in Question 13 suggests that it is not aligned with the measured main construct. Consequently, further evaluation is needed to review its wording, theoretical relevance, and suitability for the target population. According to Azwar (2019) and Nunnally & Bernstein (1994), questions deemed irrelevant should either be revised or removed after assessing their impact on the instrument's reliability. An instrument with good construct validity ensures accurate, relevant, and reliable measurements for evaluation or research purposes.

Question Reliability

Reliability measures the consistency and stability of the results of a measurement instrument when applied to the same individual on various occasions. The reliability of the test questions is assessed using Cronbach's Alpha coefficient, which ideally is above 0.7 to indicate good internal consistency (Nunnally & Bernstein (1994).

Table 3. Reliability statistics

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.763	.786	15

Table 4. Reliability statistics if item deleted

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Q1	10.99	7.100	.521	.	.737
Q2	10.95	7.524	.360	.	.751
Q3	10.97	7.299	.442	.	.744
Q4	10.99	7.009	.570	.	.732
Q5	11.00	6.818	.658	.	.724
Q6	11.14	7.324	.294	.	.759
Q7	10.99	7.191	.473	.	.741
Q8	10.89	7.570	.460	.	.747
Q9	11.02	6.981	.545	.	.734
Q10	11.24	7.568	.176	.	.773
Q11	10.89	7.715	.354	.	.753
Q12	11.10	7.563	.213	.	.766
Q13	11.18	8.022	.015	.	.788
Q14	11.04	7.399	.322	.	.755
Q15	11.08	7.130	.412	.	.746

The reliability test results on the Grade VI Mathematics Mid-Term Exam Questions of Sekolah Prestasi Global showed a Cronbach's Alpha value of 0.763, as shown in Table 3, indicating high reliability. However, if Question 13 is deleted, the Cronbach's Alpha value increases to 0.788, indicating that Question 13 has a small contribution to the instrument's consistency. This indicates that Question 13 is not optimally relevant to the main construct and can reduce the overall quality of the scale.

Further evaluation of Question 13 is needed to improve the quality of the instrument. Recommended steps include content and wording evaluation to ensure conformity to the main construct, retesting in a different population to identify sources of problems, and revision or replacement if Question 13 proves to be theoretically irrelevant or difficult to understand. Thus, removing or revising Question 13 is expected to improve the overall reliability and validity of the instrument, resulting in a more accurate and credible measurement in the educational context.

Question Parameters

In classical test theory (CTT), "question parameters" refer to the attributes used to assess and analyze test questions. The two main parameters include the number of specific questions about the question for the test taker. In

contrast, "question discrimination" refers to the ability of the question to distinguish between low-ability and high-ability test takers (LeBeau et al., 2020). Understanding these two parameters is essential to ensure the test can provide accurate and relevant information about the test taker's ability.

a. Question Discrimination Score Category

The discriminatory power of test questions refers to the test's ability to differentiate between participants with high and low abilities on the tested material (Azwar, 2019). Test questions with good discriminatory power can identify participants who understand the material in depth, thus ensuring the validity of the test in reflecting differences in participant abilities (Cappelleri et al., 2014). The discriminatory power index is calculated by comparing the proportion of correct answers between the upper group (high-achieving participants) and the lower group (low-achieving participants) using the formula (Azwar, 2019; Crocker & Algina, 2008).

$$D = \frac{P_A - P_B}{N} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

P_A is the proportion of correct answers in the upper group, P_B is the proportion of the lower group, and N is the number of participants in each group.

The discrimination index is categorized as very good ($D \geq 0.4$), good ($0.3 \leq D < 0.4$), sufficient ($0.2 \leq D < 0.3$), and poor ($D < 0.2$). Questions with low or negative discrimination indicate that the questions are ineffective and need to be revised or replaced. High discrimination not only reflects the effectiveness of the questions in measuring differences in participant abilities but also ensures that the test results are relevant and fair in accurately assessing student abilities (Azwar, 2019; Crocker & Algina, 2008).

Table 5. Level of question differential power with manual calculation

Question	A	PA	B	PB	D=PA-PB	Different power categories
Q1	30	1,000	15	0.500	0.500	Good
Q2	30	1,000	20	0.667	0.333	Enough
Q3	30	1,000	17	0.567	0.433	Good
Q4	30	1,000	15	0.500	0.500	Good
Q5	30	1,000	11	0.367	0.633	Very good
Q6	29	0.967	12	0.400	0.567	Very good
Q7	30	1,000	16	0.533	0.467	Good
Q8	30	1,000	23	0.767	0.233	Enough
Q9	30	1,000	13	0.433	0.567	Very good
Q10	28	0.933	12	0.400	0.533	Good
Q11	30	1,000	24	0.800	0.200	Enough
Q12	29	0.967	15	0.500	0.467	Good
Q13	27	0.900	19	0.633	0.267	Enough
Q14	28	0.933	14	0.467	0.467	Good
Q15	28	0.933	10	0.333	0.600	Very good

Table 5. shows the results of the discrimination power analysis show that Question 5, Question 6, Question 9, and Question 15 have very good discrimination power ($D \geq 0.60$), while Question 1, Question 3, Question 4, Question 7, Question 10, Question 12, and Question 14 are in the good category ($D \geq 0.40$), which means that these questions are effective in differentiating participants' abilities. Four questions, namely Question 2, Question 8,

Question 11, and Question 13, have sufficient discrimination power ($0.20 \leq D < 0.40$) and require revision to be more effective, for example, by strengthening distractors. There are no questions with poor discrimination power ($D < 0.20$), so all questions are still suitable for use. Overall, 10 of the 15 questions have good discrimination power, indicating a fairly good test quality, although improvements to several questions are still needed to improve measurement effectiveness.

Table 6. Question differential power level with JMetric calculation

Question	Different Power	A	B	C	D
Question1	0.9218	0.2046	0.0542	0.9218	NaN
Question2	0.9316	0.1692	0.0534	NaN	0.9316
Question3	0.9334	0.9334	0.1834	0.0417	NaN
Question4	0.9442	0.1721	0.0213	0.9442	0.0048
Question5	0.9442	0.9442	0.1664	0.0213	NaN
Question6	0.8437	0.3430	0.0624	0.0839	0.8437
Question7	0.9291	0.1609	0.0048	0.9291	0.1293
Question8	0.9612	0.0899	0.9612	NaN	NaN
Question9	0.9198	0.9198	0.2250	0.0417	NaN
Question10	0.7595	0.4226	0.1661	0.7595	0.1233
Question	**Different Power**	**B**	**S**		
Question11	0.9532	0.1133	0.9532		
Question12	0.8544	0.8544	0.3179		
Question13	0.7577	0.7577	0.4579		
Question14	0.8735	0.8735	0.3074		
Question15	0.8780	0.3029	0.8780		

Based on the analysis with JMetrik, the differentiating power of each question is shown in Table 6. The results of the discriminant power analysis using JMetric showed that 13 of the 15 questions (Questions 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 14, and 15) had very good discriminant power ($D \geq 0.80$), reflecting a high ability to differentiate participants with high and low abilities. The other two questions, namely Question 10 and Question 13, were in a good category ($0.40 \leq D < 0.80$) and only needed minimal improvement to improve their quality. Overall, the quality of the questions was considered very good. However, some distractors needed to be reviewed to ensure their effectiveness in diverting the attention of participants who did not understand the material. This finding confirms that high discriminant power in questions indicates good instrument quality. This supports the theory proposed in Classical Test Theory (CTT), where questions with high discriminant power evaluate participants' abilities more accurately.

The results of the comparative analysis of the discriminant power between manual calculations and JMetric show consistency in categorizing test questions. Questions with good to very good discriminant power ($D \geq 0.40$ D) identified manually were also confirmed by JMetric. However, JMetric excels by providing additional details, such as the specific Contribution of each answer option (A, B, C, D), which are not available in manual calculations.

For questions with very good discriminability ($D \geq 0.60$), both methods agreed on Questions 5, 6, 9, and 15. Questions with good discriminability ($0.40 \leq D < 0.6$) were also identified consistently, such as Questions 1, 3, 4, 7, 10, 12, and 14, with JMetric providing additional insight into distractor effectiveness. On questions with sufficient discriminability ($0.20 \leq D < 0.4$), such as Questions 2, 8, 11, and 13, JMetric highlights ineffective distractors, including options with NaN discriminability values that participants did not choose.

Although both methods provide consistent results, JMetric offers a more comprehensive analysis with additional details that make it easier to identify and improve test questions. Therefore, JMetric is recommended for use in question analysis to improve the overall quality of the test.

b. Question Difficulty Level

The test questions' difficulty level is a crucial aspect of test development because it affects the test's ability to accurately measure participants' abilities (Haw et al., 2022). Questions with a moderate level of difficulty ($0.31 \leq$ difficulty ≤ 0.70) are considered ideal because they can provide representative and informative variations in results in differentiating participants' abilities (Azwar 2019; Crocker & Algina, 2008). Conversely, questions that are too easy (difficulty > 0.70) or too difficult (difficulty < 0.30) tend to reduce the effectiveness of the test. In Classical Test Theory (CTT), the level of difficulty is calculated based on the proportion of participants who answer the question correctly, with moderate questions providing optimal information to evaluate differences in participant abilities.

A balanced distribution of difficulty levels is key to ensuring test quality, as it allows for more accurate and comprehensive measurements. Thus, understanding and implementing appropriate difficulty levels is essential to improving the quality and effectiveness of evaluation instruments.

The method for calculating the level of difficulty of question questions can be written as follows:

$$p_{i=\frac{\Sigma B}{N}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Information:
pi: Difficulty level of question i
B: The number of test participants who answered the questions correctly
N: Number of test participants who answered the questions.

The difficulty level index of the test questions is analyzed using a certain formula and classified into five categories based on the difficulty level value (TK). Test questions with TK = 0.00 are considered very difficult or too difficult, $0.00 < TK \leq 0.30$ are classified as difficult, $0.30 < TK \leq 0.70$ are in the moderate category, $0.70 < TK < 1.0$ are classified as easy, and TK = 1.00 are considered very easy or too easy (Sundayana, 2014). This classification helps evaluate the quality of the questions and reflects the appropriate difficulty level for the test takers.

Results of the Analysis of the Level of Difficulty of Questions from the Odd Semester Mid-Term Mathematics Exam for Grade VI Academic Year 2024 Sekolah Prestasi Global, with 15 questions sourced from 111 students using the number formula shown in Table 7.

Table 7. Difficulty level with manual calculations

Question	Difficulty Level	Difficulty Category
Q1	0.8288	Easy
Q 2	0.8739	Easy
Q 3	0.8468	Easy
Q 4	0.8288	Easy
Q 5	0.8198	Easy
Q 6	0.6757	Ideal/Moderate
Q 7	0.8288	Easy
Q 8	0.9279	Easy
Q 9	0.8018	Easy
Q 10	0.5766	Ideal/Moderate
Q 11	0.9279	Easy
Q 12	0.7207	Easy

Q 13	0.6396	Ideal/Moderate
Q 14	0.7838	Easy
Q 15	0.7387	Easy

Analysis of Table 7 shows that most questions are easy, with 12 out of 15 questions having a difficulty level above 0.71, indicating that most participants can answer them correctly. Only three questions (Question 6, Question 10, Question 13) are in the ideal category ($0.31 \leq$ difficulty ≤ 0.70), which is balanced in difficulty and suitable for evaluating participants' abilities. There are no questions with a very difficult level of difficulty (difficulty < 0.31), indicating that the overall questions tend to be too easy and not challenging enough to distinguish participants' abilities effectively.

As a comparison, the results of the analysis of the level of difficulty of the questions will be carried out using geometric software, and the results obtained are presented in Table 8 as follows:

Table 8. Question difficulty level with JMetrik calculation

Question	**Difficulty Question**	**A**	**B**	**C**	**D**	**Total**	**% Question Difficulty**	**Category**
Question1	0.0972	0.0192	0.0011	0.0972	0.0000	0.1175	0.8272	Easy
Question2	0.1036	0.0118	0.0021	0.0000	0.1036	0.1175	0.8817	Easy
Question3	0.1015	0.1015	0.0150	0.0021	0.0000	0.1186	0.8558	Easy
Question4	0.1004	0.0160	0.0011	0.1004	0.0011	0.1186	0.8465	Easy
Question5	0.0983	0.0983	0.0182	0.0011	0.0000	0.1176	0.8359	Easy
Question6	0.0833	0.0288	0.0011	0.0043	0.0833	0.1175	0.7089	Ideal/Moderate
Question7	0.0994	0.0139	0.0011	0.0994	0.0043	0.1187	0.8374	Easy
Question8	0.1100	0.0085	0.1100	0.0000	0.0000	0.1185	0.9283	Easy
Question9	0.0951	0.0951	0.0214	0.0021	0.0000	0.1186	0.8019	Easy
Question10	0.0705	0.0395	0.0053	0.0705	0.0021	0.1174	0.6005	Ideal/Moderate
Question	**Question Difficulty**	**B**	**S**					
Question11	0.1100	0.0085	0.1100			0.1185	0.9283	Easy
Question12	0.0887	0.0887	0.0267			0.1154	0.7686	Easy
Question13	0.0759	0.0759	0.0385			0.1144	0.6635	Ideal/Moderate
Question14	0.0908	0.0908	0.0256			0.1164	0.7801	Easy
Question15	0.0887	0.0278	0.0887			0.1165	0.7614	Easy

The calculation of question difficulty level with JMetric is slightly different from manual calculation in terms of numerical representation, but the categorization results remain consistent. Most questions, namely 12 out of 15, are in the easy category (difficulty > 0.71), indicating that most participants can answer them easily. Only three questions (Questions 6, 10, and 13) are in the medium / ideal category ($0.31 \leq$ difficulty ≤ 0.70), which is more balanced to evaluate participants' abilities effectively. No questions were found to be very difficult (<0.30). The small numerical differences between manual and JMetric calculations did not affect the categorization results, which, overall, this test tends to be too easy and cannot optimally distinguish participants' abilities.

JMetric calculation is more accurate than manual calculation because it provides more detailed results with precise decimals. In addition, JMetric offers additional benefits, such as analyzing the distribution of answer choices (A, B, C, D), which are not available in manual calculations.

Based on the analysis of the difficulty level of the Mathematics Mid-Term Exam questions at Sekolah Prestasi Global for the 2024/2025 Academic Year, most of the questions are classified as easy ($0.70 < TK < 1.00$), with 12 out of 15 questions (80%) in this category, namely questions number 1, 2, 3, 4, 5, 7, 8, 9, 11, 12, 14, and 15. Three questions (20%) are in the medium category ($0.30 < TK \leq 0.70$), namely questions number 6, 10, and 13. There are no questions included in the very difficult ($TK = 0.00$), difficult ($0.00 < TK \leq 0.30$), or very easy ($TK = 1.00$) categories. This distribution shows that the test tends to be less challenging, with most questions being relatively easy for participants.

The analysis shows that manual calculation and JMetric provide consistent results, which can be used interchangeably. However, the main weakness of this test is that the majority of the questions are too easy, so it is recommended to increase the difficulty level of the questions through revisions, such as adding stronger distractors or changing the wording of the questions to require more in-depth analysis. Questions with a medium/ideal level of difficulty, namely Questions 6, 10, and 13, should be maintained and used as a reference for developing new questions.

To improve the quality of the test, the distribution of difficulty levels needs to be rearranged with a composition of 30% easy questions ($0.71 \leq$ difficulty ≤ 1.0), 40% medium/ideal questions ($0.31 \leq$ difficulty ≤ 0.70) and 30% difficult questions (difficulty < 0.30). After the revision, a retest is needed to ensure that the distribution of difficulty levels is even and supports the overall effectiveness of the test.

c. Effectiveness of distractors

Distractor effectiveness is an important aspect of multiple-choice questions, as it ensures that incorrect answers attract low-comprehension participants without misleading those who understand the material well. Distractor effectiveness analysis assesses the extent to which incorrect answer choices distinguish those who understand the material from those who do not (Cappelleri et al., 2014; Suseno, 2017). In Classical Test Theory (CTT), distractors are considered effective if they can attract the attention of participants who do not understand the material. In contrast, rarely or never selected distractors are considered ineffective and must be replaced (Crocker & Algina, 2008). Thus, distractor analysis is an important part of evaluating the quality of multiple-choice questions to ensure accurate and meaningful measurements.

The analysis of distractor effectiveness in Classical Test Theory (CTT) involves several important steps. First, the proportion of participants who choose each distractor, where effective distractors are usually chosen by 5-15% of participants, especially those with low comprehension (Anastasi & Urbina, 1997). Second, the distribution of distractors should be examined to ensure that each distractor attracts participants' attention, with an even distribution of responses. Distractors that are rarely or never chosen are considered ineffective (Ebel & Frisbie, 1991). Third, distractor selectors are evaluated based on total scores, where effective distractors attract more low-scoring participants than high-scoring participants (Haladyna & Downing, 1989). If high-scoring participants choose a distractor, this indicates a problem, such as ambiguity or error in the distractor. Finally, ineffective distractors must be revised to make them more relevant or reflect common errors that participants often make (Allen & Yen, 1979). These steps ensure that the distractor functions optimally to support the quality of the multiple-choice question.

To evaluate the effectiveness of distractors manually, the following steps are taken: for example, a question has four answer options (A, B, C, D). The distribution of answers for the upper group is A (15 participants), B (correct answers, 10 participants), C (2 participants), and D (3 participants). For the lower group, the distribution is A (25

participants), B (correct answers, 5 participants), C (8 participants), and D (2 participants).

Analyzing test items begins with calculating the proportion of selections for each option. For instance, in option A, the upper group proportion is PA=$\frac{15}{30}$=0.50, while the lower group proportion is PB=$\frac{25}{30}$=00.83. Following this, each option's differential power (D) is determined by subtracting the lower group proportion from the upper group proportion (D=PA−PB). For option A, the differential power is calculated as D=0.50−0.83=−0.33. This result indicates that option A functions effectively as a distractor, as it appeals more to the lower group than the upper group. In contrast, option B, identified as the correct answer, has a differential power of D=0.33−0.17=0.16, demonstrating its effectiveness in distinguishing between high- and low-performing students. Similarly, option C yields a differential power of D=0.07−0.27=−0.20, classifying it as an effective distractor. However, option D shows a differential power of D=0.10−0.07=0.03, which is considered ineffective due to its value being close to zero, indicating poor differentiation capability.

These results indicate that options A and C function as effective distractors because they are more appealing to participants with low comprehension. In contrast, option D needs to be improved because it does not function optimally as a distractor.

Analysis using JMetric software shows variations in distractor effectiveness based on discriminant value. Options with positive discriminant values show high effectiveness in attracting low-ability participants, such as in Question 6, where option D has a discriminant value of 0.8437, indicating optimal performance as a distractor. In contrast, participants were considered ineffective and did not select options with zero or NaN discriminant value, such as in Question 8 (options C and D). Some distractors also have very low discriminant values, such as in Question 3 (option D, 0.0417) and Question 7 (option C, 0.0048), indicating minimal effectiveness.

Effective distractors, such as Question 6 (option D) and Question 9 (option B, 0.2250), can be used as references for question development. Conversely, distractors that do not function optimally must be revised to attract the attention of participants with low abilities. Retesting after revision is recommended to ensure the effectiveness of distractors is improved so that the test quality can continue to improve.

Table 9. The difference in effectiveness of distractors: Manual vs JMetric software

Aspect	Manual Calculation	JMetric
Speed	Slow, depending on the number of questions and participants	Very fast, even for large datasets
Accuracy	Prone to miscalculation	Very accurate (computer algorithm-based)
Analysis Details	Limited (only total power difference value)	Very detailed (per answer option)
Distractor Detection	Less effective for detecting rarely selected distractors	Easy to detect ineffective distractors
Data Visualization	Not available	Available in table and graph form

From the discussion above, Table 9 below summarizes the effectiveness comparison of the distractors: manual vs. JMetric Software. However, this method is less efficient for large datasets and is prone to errors. In contrast, JMetric is more effective for large datasets

because it is fast, accurate, and able to provide detailed analysis, including the contribution of each distractor and the detection of non-functioning distractors. It is recommended to use manual calculation for basic learning or simple analysis. At the same time, JMetric is more suitable for large-scale analysis or advanced evaluation due to its efficiency and high level of detail. Both methods can be complementary based on the needs and scale of the analysis.

The findings of this study reveal significant insights into the effectiveness of Grade VI Mathematics Midterm Exam questions at Sekolah Prestasi Global. Integrating Classical Test Theory (CTT) with modern analytical tools, such as SPSS and JMetric, has provided a robust evaluation framework, enabling a detailed assessment of validity, reliability, difficulty levels, discrimination power, and distractor effectiveness. This section interprets the results while connecting them to prior studies to highlight contributions and implications for test improvement.

Validity and Reliability in Context

The results demonstrated that 14 out of 15 questions were valid based on correlation values exceeding R Table (0.1865) and significance levels 0.000, confirming their alignment with the intended constructs. However, Question 13 emerged as an exception, exhibiting a correlation value of 0.182 and a significance level of 0.056. This suggests misalignment with the test's conceptual framework and necessitates revision or replacement.

The reliability analysis yielded a Cronbach's Alpha value of 0.763, indicating good internal consistency, which improved to 0.786 upon removing Question 13. These findings are consistent with prior studies, such as those by Saputri et al. (2023) and Anshari et al. 2024), emphasizing that reliability above 0.70 ensures consistent measurements. The removal of invalid questions is a key strategy in refining test quality, as corroborated by Azwar (2019) and Nunnally & Bernstein (1994).

Difficulty Levels and Discrimination Power

Analysis of difficulty levels indicated that 80% of the questions were classified as easy (difficulty > 0.70), with only three questions (Questions 6, 10, and 13) exhibiting moderate difficulty (0.31–0.70). No questions were categorized as difficult (<0.30), raising concerns about the test's ability to challenge students and distinguish between varying ability levels. This aligns with the findings of Rasmuin and Luddin (2022), who advocated for balanced distributions to improve measurement precision.

Discrimination analysis revealed that 10 of the 15 questions demonstrated good to very good discrimination power ($D \geq 0.40$), while four questions (2, 8, 11, and 13) exhibited sufficient discrimination ($0.20 \leq D < 0.40$). These results underscore the need for revisions to enhance discrimination, especially for low-performing questions. Similar conclusions were drawn by Kaldaras et al. (2024) and Suprapto et al. (2020), who emphasized integrating advanced analytical tools and frameworks to refine assessment instruments.

Effectiveness of Distractors

The effectiveness of distractors was a critical aspect of this study, as ineffective distractors reduce the overall validity and reliability of multiple-choice tests. The findings revealed weaknesses in distractor design, with several distractors exhibiting near-zero or NaN discrimination values. For instance, Question 8 contained options that failed to attract low-ability participants, rendering them ineffective.

These results parallel findings by Haladyna & Downing (1989), who stressed that distractors should target misconceptions rather than confuse high-ability students. The comparative analysis between manual calculations and JMetric reinforced the value of automated tools, particularly JMetric, which provided precise insights into distractor performance. This supports prior research, such as that by (Ghozali, 2018) and LeBeau et al.

(2020), highlighting the utility of software in large-scale assessments.

Bridging Classical and Modern Approaches

This study's manual and automated methods address the gaps in earlier research, which often relied solely on traditional approaches. The inclusion of JMetric offered deeper insights into distractor analysis and discrimination metrics, enabling a more comprehensive evaluation process. These findings complement the work by Kaldaras et al. (2024), who emphasized leveraging artificial intelligence and software tools for data-driven improvements in STEM education assessments.

Moreover, this research builds on the methodologies proposed by Crocker & Algina (2008) and Anastasi and Urbina (1997) while adapting modern statistical tools to traditional test theories. This study enhances test evaluation practices by bridging technology with established frameworks, offering scalable solutions for diverse educational settings.

The results underscore the need for revising easy questions to introduce greater cognitive challenges and improve differentiation among students. Recommendations include:

1. Balanced Distribution of Difficulty Levels: Adjust the question composition to 30% easy, 40% moderate, and 30% difficult to ensure a fair evaluation framework (Azwar, 2019)
2. Improvement of Distractors: Replace ineffective distractors with plausible alternatives that reflect common misconceptions (Haladyna & Downing, 1989)
3. Retesting After Revisions: Conduct pilot testing post-revision to validate improvements and assess reliability and discrimination consistency (Nunnally & Bernstein, 1994)
4. Incorporation of Software Tools: Use JMetric and similar platforms for large-scale question analysis, (LeBeau et al., 2020).

These recommendations resonate with prior findings by (Iskandar & Rizal, 2018) and (Susdelina et al., 2018), who emphasized iterative testing and evaluation cycles to maintain assessment quality. Additionally, adopting frameworks proposed by (Zakiyah and Kartika, 2024) for mathematical instruments could enhance question design and development.

The findings from this study align closely with prior research on question quality, validity, and reliability while contributing novel insights into distractor effectiveness through a multi-method approach. By harmonizing traditional test theories with modern analytical tools, this study provides a framework for improving educational assessments that are more valid, reliable, and capable of accurately measuring student abilities. Future research could explore integrating AI-driven assessment tools, as proposed by (Kaldaras et al., 2024), to further enhance test quality and adaptability.

IV. Conclusion

This study shows that the Mathematics Mid-Term Exam questions for grade VI at Sekolah Prestasi Global are of quite good quality based on the analysis of classical test theory. The reliability of the test is considered good, with a Cronbach's Alpha of 0.763, which increased to 0.788 after removing Question 13, indicating that this question is less relevant to the main construct. The difficulty and discrimination of the questions are generally in the moderate to good category, but some distractors were found to be less effective. Most of the questions are classified as easy, with 13 out of 15 questions having a difficulty level above 0.70. Only two questions are in the ideal category. At the same time, there are no questions with a high level of difficulty, so they are less able to provide variation to evaluate participants' abilities. Distractors on most questions are also considered ineffective, indicating the need for improvement in the distribution of difficulty levels and the effectiveness of the questions.

The results of the comparison of manual analysis and using JMetric show consistency in the grouping of discrimination power and difficulty levels of the test questions. JMetric excels in providing more detailed analysis, including detecting distractors' effectiveness,

although most distractors do not function optimally. As many as 12 of the 15 questions are classified as easy, only three are at the ideal difficulty level, and no questions are too difficult, indicating that this test is not challenging enough to differentiate participants' competency level significantly. Although it has good quality, this test needs revision to create a more balanced distribution of difficulty levels and improve evaluation accuracy.

This study recommends revising invalid or ineffective questions, such as Question 13, and improving less interesting distractors. Questions with difficulty levels that are too easy (>0.70), such as Questions 1, 2, and 3, need to be increased in difficulty to distinguish participants with different abilities. In contrast, questions with ideal difficulty levels, such as Questions 10 and 13, can be maintained as references for development. Question revision involves adding analysis elements and stronger distractors and arranging the proportion of difficulty levels to 30% easy (0.71–1.00), 40% medium (0.31–0.70), and 30% difficult (0.00–0.30). Improvement of distractors is also done by replacing weak or unrealistic distractors to be more effective. After revision, retesting is needed to ensure the validity and reliability of the questions involving a wider population. JMetric is recommended for large dataset analysis, although manual calculations are still relevant for basic teaching. The results of the analysis can be used as a reference for developing questions in the future to have an optimal level of difficulty and discrimination power, as well as more effective distractors.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, Calif. : Brooks/Cole Pub. Co.

Anastasi, A. , & Urbina, S. (1997). *1 Anastasia* ((7th ed.)). Prentice Hall/Pearson Education.

Anshari, M. I., Nasution, R., Irsyad, M., Alifa, A. Z., & Zuhriyah, I. A. (2024). Analisis validitas dan reliabilitas butir soal sumatif akhir semester ganjil mata pelajaran PAI. *EDUKATIF : Jurnal Ilmu Pendidikan*, *6*(1), 964–975. https://doi.org/10.31004/edukatif.v6i1.5931

Azwar, S. (2019). *Reliabilitas dan validitas* (IV). Pustaka Pelajar.

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Crocker, L. M. ., & Algina, James. (2008). *Introduction to classical and modern test theory* (M. S. Michele Baird Maureen Staudt, Ed.). Cengage Learning.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th Edition). Englewood Cliffs, N.J. : Prentice-Hall.

Ghozali, I. (2018). *Aplikasi analisis multivariate dengan program IBM SPSS 25 edisi ke-9* (ke-9). Universitas Diponegoro.

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 51–78. https://doi.org/10.1207/s15324818ame0201_4

Haw, L. H., Sharif, S. B., & Han, C. G. K. (2022). Analyzing the science achievement test: Perspective of classical test theory and Rasch analysis. *International Journal of Evaluation and Research in Education*, *11*(4), 1714–1724. https://doi.org/10.11591/ijere.v11i4.22304

İlhan, M. (2016). A Comparison of the results of many-facet Rasch analyses based on crossed and judge pair designs. *Educational Sciences: Theory and Practice*, *16*(2), 579-601.

Iskandar, A., & Rizal, M. (2018). Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(1), 12–23. https://doi.org/10.21831/pep.v22i1.15609

Kaldaras, L., Haudek, K., & Krajcik, J. (2024). Employing automatic analysis tools aligned to learning progressions to assess knowledge application and support learning in STEM. In *International Journal of STEM Education* (Vol. 11, Issue 1). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1186/s40594-024-00516-0

LeBeau, B., Assouline, S. G., Mahatmya, D., & Lupkowski-Shoplik, A. (2020). Differentiating among high-achieving learners: A comparison of classical test theory and item response theory on above-level testing. *Gifted Child Quarterly*, *64*(3), 219–237. https://doi.org/10.1177/0016986220924050

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. (3rd ed.). McGraw-Hill, Inc., New York.

Rasmuin, R., & Luddin, S. (2022). Tingkat kesulitan soal buatan guru bidang studi matematika menurut teori tes klasik pada tingkat SMP di Kota Baubau. *Jurnal Akademik Pendidikan Matematika*, 33–40. https://doi.org/10.55340/japm.v8i1.699

Saputri, H. A., Zulhijrah, Larasati, N. J., & Saleh. (2023). Analisis instrumen assesmen validitas, reliabilitas, tingkat kesukaran dan daya beda butir soal. 9. https://doi.org/ttps://doi.org/10.36989/didaktik.v9i5.2268

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch: Pada assessment pendidikan* (cetakan 1). Trim Komunikata.

Sundayana, R. (2014). *Educational research statistics*. Alphabet.

Suprapto, E., Saryanto, S., Sumiharsono, R., & Ramadhan, S. (2020). The analysis of instrument quality to measure the students' higher order thinking skill in physics learning. *Journal of Turkish Science Education*, *17*(4), 520–527. https://doi.org/10.36681/tused.2020.42

Susdelina, Perdana, S. A., & Febrian. (2018). Analisis kualitas instrumen pengukuran pemahaman konsep persamaan kuadrat melalui teori tes klasik dan Rasch model. *Jurnal KIPRAH*, *VI*(1) 41–48. https://doi.org/https://doi.org/10.31629/kiprah.v6i1.574

Suseno, I. (2017). Komparasi karakteristik butir tes pilihan ganda ditinjau dari teori tes klasik. *Faktor Jurnal Ilmiah Kependidikan*, *4*(1), 1–8. https://doi.org/http://dx.doi.org/10.30998/fjik.v4i1.1588

Zakiyah, Z., & Kartika, H. (2024). Uji validitas konten instrumen kemampuan representasi matematis dalam menyelesaikan masalah bangun datar. *JP2M (Jurnal Pendidikan Dan Pembelajaran Matematika)*, *10*(1), 250–257. https://doi.org/10.29100/jp2m.v10i1.5485