Analysis of Instrument Quality in Measuring Mathematical Logical Thinking Ability

Yulia Pratiwi[1], Niki Pujarwati[2], Marlinawati[3], Agung Putra Wijaya[4*)], Chika Rahayu[5]
[1,2,3,4,5]Lampung University
[*)] agung.wijaya@fkip.unila.ac.id

Abstract

The ability to think logically in mathematics is a skill that must be developed through instruction. One strategy for developing this ability is the use of high-quality instruments. Therefore, this study aims to analyze the quality of a mathematical logical thinking ability test instrument in terms of its validity, reliability, difficulty index, and discriminating power using SPSS and Microsoft Excel. This research employed a quantitative method with an evaluative analysis approach. The subjects in this study were students of class VII.6 and VII.8 at SMPN 26 Bandarlampung. The instrument used was a descriptive test consisting of ten items, developed based on four indicators of logical thinking ability, namely the ability to interpret information, the ability to predict, the ability to solve problems, and the ability to draw conclusions. The results of this study indicate that: (1) the quality of the logical thinking ability test instrument meets the criteria for 100% validity; (2) the reliability of the logical thinking ability test instrument is 0.775, categorized as reliable; (3) the difficulty level analysis shows that nine items fall into the moderate category and one item falls into the easy category; (4) the discrimination power analysis shows that four items have good discriminating power and six items have fair discriminating power.

Keywords: Mathematical logical thinking ability, test instrument, validity, reliability, difficulty index, discriminating power

Introduction

Education in the 21st century demands changes in the learning process because students can no longer be equipped merely with memorization skills; they need to be trained to master higher-order thinking skills that are relevant to contemporary developments. OECD (2021) emphasizes that education must help students adapt to very rapid changes by developing a variety of essential skills, rather than simply recalling information, so that they are able to face the challenges of modern technology. In line with this, the *Merdeka* curriculum in Indonesia explicitly prioritizes the strengthening of HOTS (Higher Order Thinking Skills) through deep learning, student-centered instruction, and the promotion of critical, logical, and creative thinking, as well as problem-solving skills (Kemendikbudristek, 2025). The importance of integrating these skills into the curriculum is reinforced by recent research (Chusna et al., 2024), which affirms that 21st-century skills are crucial for preparing the younger generation for the era of Society 5.0, in which humans' ability to think, analyze, and innovate becomes a key factor that cannot be replaced by automated systems.

Logical thinking ability is an essential component of mathematics learning, as it enables students to reason systematically, understand relationships among concepts, and draw valid conclusions, thereby supporting deep conceptual understanding and effective problem solving (Hadi et al., 2025). Nevertheless, international assessment results indicate that students' reasoning abilities remain inadequate. The PISA 2022 findings reveal that Indonesian students' mathematics achievement is still below the OECD average, particularly on tasks requiring logical reasoning and inference (OECD, 2023). This condition underscores the need for assessment practices that go beyond procedural mastery and are capable of accurately measuring students' logical thinking through well-designed instruments. Accordingly, this study defines mathematical logical thinking ability as students' capacity to process mathematical information systematically, as reflected in four indicators: interpreting information by identifying relevant data and relationships, predicting by generating reasonable conjectures based on given conditions, solving problems through the application of appropriate strategies and logical reasoning steps, and drawing coherent and valid conclusions from the obtained results. These indicators form the basis for the development of test items and scoring criteria.

Given the low level of students' logical thinking ability as indicated by the PISA results, learning evaluation needs to be directed toward truly measuring aspects of reasoning, which are the main components in learning mathematics. Evaluation essentially functions to identify the extent to which students are able to use logic, understand relationships between concepts, and draw valid conclusions, so that assessment instruments must be designed in accordance with the characteristics of the abilities to be measured (Brookhart, 2018). A quality assessment instrument is not sufficient if it only tests factual knowledge; it must be designed to measure higher-order thinking skills, such as logical reasoning in mathematics, through items that demand in-depth analysis, conclusion drawing, and sound decision-making (Popham, 2020). In this context, item quality becomes an important factor because items that meet the criteria of validity, reliability, difficulty index, and discriminating power are able to measure the intended ability accurately (Arifin, 2022). Conversely, weak items can produce biased or misleading information about students' level of mastery, thus leading to inappropriate instructional decisions. Therefore, the measurement of logical thinking ability is highly dependent on the quality of the evaluation instruments used, which must meet the principles of validity, reliability, difficulty index, and discriminating power.

In various mathematics studies, it has been found that most of the test items used in school examinations have not been thoroughly analyzed in terms of their quality. Important aspects such as validity, reliability, difficulty index, and discriminating power are often not examined or are overlooked (Nafs et al., 2023). Issues such as invalid items, low reliability, imbalanced difficulty indexs, and low discriminating power indicate that the mathematics test items in circulation do not meet the quality standards of an assessment instrument.

This condition affirms that item analysis is a crucial step in evaluating the quality of each test component, including validity, reliability, difficulty index, and discriminating power (Manfaat et al., 2021). Without such analysis, test developers do not obtain empirical information regarding item quality, so invalid or unreliable items may still be used in evaluation (Kurniati et al., 2025). Item analysis should ideally be conducted before the test is administered so that decisions made are based on valid and accurate data. Without this step, the interpretation of test results can be misleading and may not accurately reflect students' actual logical thinking abilities (Ridwan et al., 2021). In addition, item analysis provides long-term benefits for improving evaluation instruments. Teachers can identify which items need to be revised, replaced, or refined, such as items with low discriminating power or items that are too easy or too difficult.

Although the evaluation of mathematical ability, including logical thinking skills, is frequently carried out in schools, comprehensive item analysis is still rarely implemented. This is in line with the study conducted by Hakim and Revita (2025), which found that although the essay items were declared valid, the overall reliability of the instrument remained low, and the discriminating power of several items also varied, meaning that not all items were able to clearly distinguish between high- and low-ability students. Meanwhile, studies that conduct a thorough analysis including aspects such as validity, reliability, difficulty index, and discriminating power are still relatively limited in the context of mathematical logical thinking items at the secondary school level.

At present, there is still very little research in Indonesia that specifically analyzes the quality of test items designed to measure logical thinking ability. Studies that examine students' logical thinking or mathematical reasoning generally focus on developing test items or improving students' abilities through instructional models, rather than on analyzing the quality of the items used to measure those abilities. This is in line with the study by Nursyahidah et al. (2016), which measured students' reasoning, but did not conduct item analysis to determine whether the items truly measured reasoning. The

scarcity of studies that combine item analysis with indicators of logical thinking shows the existence of a research gap that needs to be addressed, particularly in order to produce mathematics evaluation instruments that are genuinely capable of accurately measuring students' logical thinking abilities.

Given these conditions, this study is important to conduct because the quality of assessment instruments plays a crucial role in ensuring that the learning evaluation process truly measures the abilities that are intended to be developed, particularly students' logical thinking skills. Instruments that are not analyzed in depth have the potential to produce inaccurate information about students' abilities, which in turn affects instructional decision-making in the classroom. Therefore, this study provides benefits not only for researchers, but also for teachers and schools in providing evaluation instruments that are more valid, reliable, and aligned with curriculum demands. Teachers can use instruments whose quality has been tested to assess logical thinking skills more accurately, while schools can improve the quality of their assessment practices through the use of more targeted measurement tools. In addition, the results of this study are expected to contribute to the development of more comprehensive mathematics evaluation instruments in future research. Based on this urgency, this study generally aims to analyze the quality of test items specifically designed to measure logical thinking ability so that the instruments used are truly valid and fit for purpose.

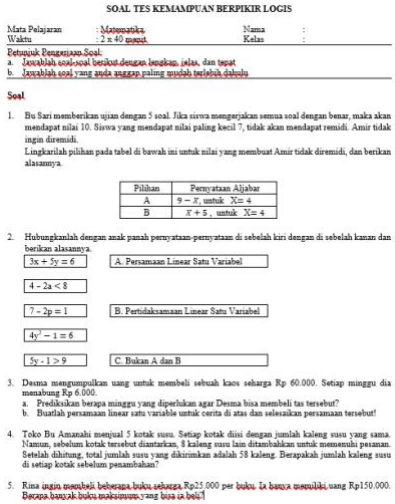
Method

This research is a quantitative study with an evaluative analysis approach, focusing on examining the quality of a test instrument measuring mathematical logical thinking ability. According to Susanto et al. (2015) and Arifin (2022), evaluative analysis in educational measurement is conducted to assess the quality and effectiveness of test items as measurement tools through indicators of validity, reliability, item difficulty, and discrimination power, thereby ensuring that the instrument truly represents the construct being measured. Therefore, this approach was adopted to ensure that each item in the logical thinking ability test functions appropriately as a measurement instrument. The participants in this study were students from two Grade VII classes (VII.6 and VII.8) at a public junior high school in Bandar Lampung. The total number of participants was 64 students, consisting of 32 students from class VII.6 and 32 students from class VII.8. A purposive sampling technique was employed, with class selection based on similarities in

academic characteristics and the completion of relevant mathematics content. Both classes were used as trial groups to evaluate the quality of the test instrument.

The instrument used was an essay test consisting of ten items, developed based on four indicators of logical thinking ability, namely the ability to interpret information, the ability to predict, the ability to solve problems, and the ability to draw conclusions. The data analysis used in this study included tests of validity, reliability, difficulty index, and discriminating power. The calculations for validity, reliability, difficulty index, and discriminating power were conducted with the aid of SPSS and Microsoft Excel.

The questions below were used to measure mathematical logical thinking ability.



SOAL TES KEMAMPUAN BERPIKIR LOGIS

Mata Pelajaran : Matematika Nama :
Waktu : 2 x 40 menit Kelas :

Petunjuk Pengerjaan Soal:
a. Jawablah soal-soal berikut dengan lengkap, jelas, dan tepat
b. Jawablah soal yang anda anggap paling mudah terlebih dahulu

Soal

1. Bu Sari memberikan ujian dengan 5 soal. Jika siswa mengerjakan semua soal dengan benar, maka akan mendapat nilai 10. Siswa yang mendapat nilai paling kecil 7, tidak akan mendapat remidi. Amir tidak ingin diremidi.
   Lingkarilah pilihan pada tabel di bawah ini untuk nilai yang membuat Amir tidak diremidi, dan berikan alasannya.

Pilihan	Pernyataan Aljabar
A	$9 - X$, untuk $X = 4$
B	$X + 5$, untuk $X = 4$

2. Hubungkanlah dengan anak panah pernyataan-pernyataan di sebelah kiri dengan di sebelah kanan dan berikan alasannya.

$3x + 5y = 6$	A. Persamaan Linear Satu Variabel
$4 - 2a < 8$	
$7 - 2p = 1$	B. Pertidaksamaan Linear Satu Variabel
$4y^2 - 1 = 6$	
$5y - 1 > 9$	C. Bukan A dan B

3. Desma mengumpulkan uang untuk membeli sebuah kaos seharga Rp. 60.000. Setiap minggu dia menabung Rp 6.000.
   a. Prediksikan berapa minggu yang diperlukan agar Desma bisa membeli tas tersebut?
   b. Buatlah persamaan linear satu variable untuk cerita di atas dan selesaikan persamaan tersebut!

4. Toko Bu Amanahi menjual 5 kotak susu. Setiap kotak diisi dengan jumlah kaleng susu yang sama. Namun, sebelum kotak tersebut diantarkan, 8 kaleng susu lain ditambahkan untuk memenuhi pesanan. Setelah dihitung, total jumlah susu yang dikirimkan adalah 58 kaleng. Berapakah jumlah kaleng susu di setiap kotak sebelum penambahan?

5. Rina ingin membeli beberapa buku seharga Rp25.000 per buku. Ia hanya memiliki uang Rp150.000. Berapa banyak buku maksimum yang bisa ia beli?

Figure 1. Logical Thinking Ability Test Questions

The interpretation of instrument validity, reliability, difficulty index, and discriminating power follows the criteria below:

1. Validity

The validity test employed in this study was criterion-related validity, which was analyzed using the Pearson product moment correlation with the assistance of SPSS. According to Ghozali (2018), an item is considered valid if it has a significance value of less than 0.05.

2. Reliability

Test reliability refers to the consistency and dependability of the measurement results. The main purpose of estimating reliability is to determine the extent to which the test scores are accurate and stable. Reliability was measured using Cronbach's alpha coefficient to assess internal consistency with the assistance of SPSS. The interpretation of the reliability coefficients according to Sudijono (2011) is shown below:

Table 1. Reliability Criteria

Reliability Coefficient (r_{11})	Interpretation
$r_{11} < 0{,}70$	Not Reliable
$r_{11} \geq 0{,}70$	Reliable

3. Difficulty Index

The formula used to calculate the item difficulty index for essay-type questions according to Arifin (2012) is as follows:

$$Difficulty\ Index = \frac{Mean\ item\ score}{Maximum\ score\ for\ each\ item}$$

The classification of the difficulty index according to Arikunto (2013) is shown below:

Table 2. Difficulty Index Criteria

Difficulty Index	Interpretation
0,00 – 0,30	Difficult
0,31 – 0,70	Moderate
0,71 – 1,00	Easy

4. Discriminating Power

Discriminating power refers to the extent to which an item can differentiate between high-performing and low-performing students (Asrul, 2015). The formula used to calculate the discrimination index of essay-type items according to Arifin (2012) is as follows:

$$DP = \frac{\bar{X}K_A + \bar{X}K_B}{Maximum\ score}$$

Where:

DP = Discrimination index

$\bar{X}K_A$ = Mean score of the upper group

$\bar{X}K_B$ = Mean score of the lower group

The classification of discriminating power based on Arikunto (2013) is as follows:

Table 3. Discriminating Power Criteria

Discriminating Power (DP)	Interpretation
$0{,}7 < DP \leq 1{,}0$	Excellent
$0{,}4 < DP \leq 0{,}7$	Good
$0{,}2 < DP \leq 0{,}4$	Fair
$0{,}0 < DP \leq 0{,}2$	Poor
$DP \leq 0{,}0$	Very Poor

Results and Discussion

The quality of the test instrument based on the aspects of validity, reliability, difficulty index, and discriminating power was obtained as follows. The type of validity

used is criterion-related validity. The criterion validity test was carried out using SPSS. Based on the analysis using SPSS, the following data were obtained.

Table 4. Results of the instrument validity analysis

Item Number	Sig. (2-tailed)	Notes
1.	0,000	Valid
2a	0,000	Valid
2b	0,000	Valid
2c	0,000	Valid
2d	0,000	Valid
2e	0,000	Valid
3a	0,000	Valid
3b	0,000	Valid
4	0,000	Valid
5	0,000	Valid

Table 5. The results of the reliability analysis using SPSS

Cronbach's Alpha	N of Item
0,775	10

A Cronbach's alpha value greater than 0.7 indicates that the test instrument is reliable.

The results of the difficulty index analysis calculated using Microsoft Excel

Table 6. Difficulty Index

Item Number	Indicator	Difficulty level	Notes
1.	Problem solving Drawing conclusion	0.56	Moderate
2a	Drawing conclusion	0.68	Moderate
2b	Drawing conclusion	0.69	Moderate
2c	Drawing conclusion	0.68	Moderate
2d	Drawing conclusion	0.71	Easy
2e	Drawing conclusion	0.67	Moderate
3a	Interpreting Predicting	0.62	Moderate
3b	Problem solving Drawing conclusion	0.50	Moderate
4	Interpreting Predicting Problem solving Drawing Conclusion	0.56	Moderate
5	Interpreting Predicting Problem solving Drawing Conclusion	0.41	Moderate

The results of the discriminating power analysis calculated using Microsoft Excel

Table 7. Discriminating Power

Item Number	Indicator	Discriminating power	Notes
1.	Problem solving Drawing conclusion	0.44	Good
2a	Drawing conclusion	0.39	Fair
2b	Drawing conclusion	0.39	Fair
2c	Drawing conclusion	0.30	Fair
2d	Drawing conclusion	0.39	Fair
2e	Drawing conclusion	0.39	Fair
3a	Interpreting Predicting	0.54	Good
3b	Problem solving Drawing conclusion	0.33	Fair
4	Interpreting Predicting Problem solving Drawing Conclusion	0.47	Good
5	Interpreting Predicting Problem solving Drawing Conclusion	0.41	Good

The try-out of the test instrument was carried out to measure junior high school students' logical thinking ability based on four indicators, namely the ability to interpret problems, to predict, to solve problems, and to draw conclusions. The instrument used was an essay-type test consisting of ten items, which was piloted with Grade VII students (classes VII.6 and VII.8) at a public junior high school in Bandar Lampung. The choice of essay format enabled the researcher to observe students' thinking processes more comprehensively, not only from their final answers but also from the way they constructed arguments and solution steps. Thus, each item did not merely represent mathematical content, but was also designed to reveal the level of logical thinking ability in accordance with the predetermined indicators.

The results of the validity analysis show that all items meet the criteria for criterion-related validity. The test, conducted using SPSS with Pearson's correlation coefficient,

indicates that each item has a positive and significant correlation with the total instrument score. Significance values below 0.05 indicate that each item contributes meaningfully to the measurement of the construct of logical thinking ability. Correlations in the moderate to high range show that these items are effective in capturing variations in students' abilities related to interpreting information, predicting possibilities, solving problems, and drawing conclusions logically.

From a measurement theory perspective, good criterion-related validity indicates that the instrument has a strong empirical basis to be used as a measuring tool. This is particularly important because logical thinking ability is an abstract construct that cannot be observed directly. A valid instrument helps ensure that the scores obtained by students truly reflect their logical thinking ability, rather than merely their ability to memorize procedures or their chance of answering correctly.

The reliability analysis carried out using Cronbach's alpha produced a coefficient of 0.775. This value is above the 0.70 threshold that is generally used as a criterion indicating that an instrument has adequate reliability. Thus, the instrument can be categorized as having good internal consistency. This means that the items in the instrument function well in measuring the same construct and do not contradict one another. If this instrument is used with different groups of students who share similar characteristics, there is a high probability that the pattern of results obtained will be relatively consistent.

Such consistency is very important in the context of measuring logical thinking ability, because this construct is influenced not only by cognitive factors but can also be affected by external factors such as emotional conditions or learning situations. Good reliability provides confidence that variations in students' scores are largely due to differences in logical thinking ability, rather than to inaccuracies in the measuring instrument. Therefore, this instrument has the potential to be used in further research as well as as an evaluation tool in classroom learning.

In terms of difficulty index, the analysis shows that nine items fall into the moderate category and one item into the easy category. This composition indicates that, in general, the instrument is at a moderate index of difficulty. A moderate difficulty index is ideal for an instrument that aims to uncover higher-order thinking skills, such as logical thinking ability, because students are still challenged to think, but the items are not so difficult as to create unnecessary barriers. The presence of one easy item can also be

considered positive as long as it remains relevant to the logical thinking indicators, since items with a low difficulty index can help build students' confidence and assist in distinguishing students with very low ability from those with moderate or high ability.

An excessively high difficulty index risks preventing many students from demonstrating their actual logical thinking ability due to excessive cognitive load. Conversely, if there are too many easy items, the instrument can no longer effectively challenge students to display more complex logical thinking. Therefore, the dominance of moderate-level items indicates that the researcher has considered a balance between cognitive demands and the accessibility of the items for students.

The analysis of item difficulty revealed variations in students' performance across the indicators of logical thinking ability. Most items were classified as having a moderate level of difficulty (P = 0.41–0.69), indicating that the instrument has a proportional difficulty level and is appropriate for use. From an indicator-based perspective, items measuring the ability to draw conclusions were generally categorized as moderate, with one item classified as easy (P = 0.71), suggesting that this indicator is relatively easier for students to master. In contrast, the predicting indicator, particularly items that integrate interpretation and problem solving, exhibited a higher level of difficulty, with the lowest value observed in Item 5 (P = 0.41), indicating that predictive activities remain challenging for students. Meanwhile, the indicators of interpreting information and problem solving were within the moderate category with relatively balanced values (P = 0.50–0.62). Overall, this analysis confirms that drawing conclusions tends to be easier, whereas predicting is more difficult for students, indicating that the developed instrument has diagnostic value in identifying indicators that require greater attention in mathematics instruction.

From the perspective of discriminating power, the instrument shows reasonably good quality. Four items fall into the "good" discrimination category, while the remaining six are in the "fair" category. The absence of items with poor or very poor discriminating power is an important indicator that all items are able, to varying degrees, to distinguish between students with high and low logical thinking ability. Items with good discriminating power function as highly effective questions for identifying students who truly have superior logical thinking skills, because such items tend to be answered correctly by high-ability students and less accurately by low-ability students.

The analysis showed that the item discrimination indices ranged from 0.30 to 0.54, indicating fair to good discrimination and suggesting that the instrument effectively differentiates students' logical thinking abilities. Items assessing interpretation and prediction demonstrated the highest discrimination power, particularly Item 3a (DP = 0.54), making them the most effective in distinguishing between high- and low-performing students. Items integrating multiple indicators also exhibited good discrimination (DP = 0.41–0.47), reflecting the sensitivity of complex reasoning tasks to ability differences. In contrast, items measuring drawing conclusions showed lower discrimination (DP = 0.30–0.39), consistent with their relatively lower difficulty levels. Overall, indicators involving prediction and higher-order reasoning demonstrated stronger discrimination power, indicating that the instrument not only meets item quality standards but also has diagnostic value for identifying variations in students' logical thinking abilities.

Meanwhile, items in the fair discrimination category can still be retained in the instrument, but they also offer room for improvement. Future researchers may revise the wording, context, or scoring scheme of such items to enhance their discriminating power—for instance by clarifying the required steps of reasoning or adding complexity to the situation within reasonable limits. Nevertheless, the current composition already indicates that the instrument functions adequately as a tool for selection and diagnosis of logical thinking ability.

When the four aspects validity, reliability, difficulty index, and discriminating power are considered together, it can be concluded that the test instrument developed has good quality and is suitable for measuring junior high school students' logical thinking ability. All items have been proven valid, the instrument as a whole is reliable, the difficulty index tends to fall in the moderate category, and the discriminating power ranges from fair to good. This combination of findings strengthens the position of the instrument as a measuring tool that is not only theoretically and empirically sound, but also operationally feasible in the field.

The findings of this study are consistent with previous research conducted by Hartono et al. (2023) and Fradinata et al. (2025), which reported that well-constructed mathematics instruments generally demonstrate acceptable levels of validity and reliability. Most of the items in this study were classified as having a moderate level of difficulty, indicating that the instrument is appropriate for measuring higher-order thinking skills. However, in contrast to several earlier studies that identified items with low discrimination

power, this study did not find any items categorized as having poor discrimination, suggesting better item quality and construction.

Substantively, this instrument provides a fairly comprehensive picture of students' logical thinking profiles, particularly in interpreting problems, predicting solutions, solving mathematical problems, and drawing conclusions coherently. Teachers can use the test results to identify which indicators are still weak in their students and to design strategies for improvement. Thus, the instrument serves not only as an evaluation tool but also as a basis for pedagogical decision-making.

The findings of this study contribute a model of an instrument that can be used as a reference for developing tools to measure logical thinking ability at different levels and in different contexts. The indicator structure used, the essay-item format, and the item analysis procedures applied can be adapted by other researchers interested in examining higher-order thinking skills, whether in mathematics or in other subject areas that demand logical reasoning.

Conclusion and Suggestion

The results of the study show that the test instrument developed to measure students' logical thinking ability has, overall, met the eligibility criteria. All items were found to be valid based on the criterion-related validity test, confirming that each item is able to reflect aspects of logical thinking ability in accordance with the specified indicators. The reliability coefficient of 0.775 also confirms that the instrument has adequate internal consistency and is capable of producing relatively stable scores when administered to groups of students with similar characteristics.

In terms of item characteristics, the majority of questions fall into the moderate difficulty category, indicating that the instrument provides a balanced level of challenge for students. In addition, the analysis of discriminating power shows that the items are able to classify students according to their level of logical thinking ability fairly well.

Based on these findings, the instrument can be declared suitable for use as a measuring tool for logical thinking ability, both in research and for instructional evaluation purposes. The instrument is effective in measuring students' ability to interpret problems, make predictions, solve problems, and draw conclusions in accordance with the indicators of logical thinking ability.

Suggestion

1. For Mathematics Teachers

The instrument analyzed in this study can serve as a reference for designing test items that assess students' logical thinking ability. Teachers are encouraged to conduct validity, reliability, difficulty level, and discrimination index analyses before using test items in evaluations to ensure that the assessment results accurately reflect students' abilities.

2. For Schools

A system is needed that encourages teachers to routinely conduct item analysis during the preparation of every evaluation to ensure the use of high-quality assessment instruments.

3. For Future Researchers

Future studies may be expanded by increasing the sample size or involving different grade levels to obtain a more comprehensive understanding of the quality of instruments used to measure logical thinking skills.

4. For Instrument Development

Several test items that fall within the "fair" discrimination category should be revised to improve their quality. Revisions may include refining the context of the questions, adjusting the level of complexity, or enhancing the clarity of indicators so that the items can more effectively distinguish between high- and low-ability students.

References

Arikunto, S. 2013. *Dasar-dasar evaluasi pendidikan* (Edisi revisi). Jakarta: Bumi Aksara.

Arifin, D. Z. (2012). *Evaluasi Pembelajaran*. Jakarta. Direktorat Jenderal Pendidikan Islam Kementrian Agama

Arifin, Z. (2022). *Evakuasi Pembelajaran: Prinsip, Teknik, dan Prosedur*. Jakarta: RajaGrafindo Persada

Asrul. 2015. *Evaluasi pembelajaran*. Bandung: Citapustaka Media.

Chusna, I. F., Aini, I. N., Putri, K. A., & Elisa, M. C. (2024). Literatur review: Urgensi keterampilan abad 21 pada peserta didik. *Jurnal Pembelajaran, Bimbingan, Dan Pengelolaan Pendidikan*, *4*(4), 1.

Fradinata, Z., Kartini, & Maimunah. (2025). Instrumen tes kemampuan representasi matematis materi relasi dan fungsi. *Jurnal Pendidik Indonesia*, *6*(2), 88–98. https://doi.org/10.61291/jpi.v6i2.98

Ghozali, I. (2018). *Aplikasi Analisis Multivariate dengan Program IBM SPSS 25*. Badan Penerbit Universitas Diponegoro.

Hadi, S., Dolk, M., Kamaliyah, & Hidayanto, T. (2025). Mathematical reasoning: How students learn mathematics? *Journal on Mathematics Education*, *16*(3), 937–954. https://doi.org/10.22342/jme.v16i3.pp937-954

Hakim, A. R., & Revita, R. (2025). Quality Analysis of Evaluation Instruments for Junior High School Students' Mathematical Conceptual Understanding. *MATH-EDU: Jurnal Ilmu Pendidikan Matematika*, *10*(2), 202–211.

Hartono, W., Hadi, S., Rosnawati, R., Retnawati, H. (2023). Exploration of Diagnostic Testing Instruments: Validity, Reliability, and Item Characteristics. *Pegem Journal of Education and Instruction*, *13*(3). https://doi.org/10.47750/pegegog.13.03.39

Kurniati, A., Rahmi, D., Yuniati, S., & Rahmania, D. (2025). *Analisis Butir Soal Tes Materi Pertidaksamaan Liner untuk Siswa Kelas XI*. *09*, 1703–1714. https://doi.org/10.31004/cendekia.v9i3.4017

Manfaat, B., Nurazizah, A., & Misri, M. A. (2021). Analysis of mathematics test items quality for high school. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *25*(1), 108–117.

Nafs, H., Sridana, N., Hikmah, N., & Soeprianto, H. (2023). Analisis kualitas butir soal ulangan akhir semester genap mata pelajaran matematika kelas vii smpn 6 mataram tahun ajaran 2022/2023. *Jurnal Ilmiah Profesi Pendidikan*, *8*(4), 2324–2331.

Nursyahidah, F., Saputro, B. A., & Prayitno, M. (2016). Kemampuan penalaran matematis siswa smp dalam belajar garis dan sudut dengan geogebra. *Suska Journal of Mathematics Education*, *2*(1), 13–19.

Ridwan, M. R., Istiyono, E., & Widihastuti, W. (2021). Test Items Analysis of Mathematical Problem Solving Ability using a Classical Test Theory Approach. *Jurnal Pendidikan MIPA*, *22*(1), 98–111.

Susanto, H., Rinaldi, A., & Islam Negeri Raden Intan Lampung, U. (2015). Analisis Validitas Reabilitas Tingkat Kesukaran dan Daya Beda pada Butir Soal Ujian Akhir Semester Ganjil Mata Pelajaran Matematika. In *Jurnal Pendidikan Matematika* (Vol. 6, Issue 2).