EMTEKA: Jurnal Pendidikan Matematika
Volume 7, No. 1, 2025. 149-161
DOI: https://doi.org/10.24127/emteka.v6i1.7474

e-ISSN 2746-5594
p-ISSN 2746-5608

ANALYSIS OF DESCRIPTIVE QUESTIONS IN DAILY TESTS FOR THE MATHEMATICS SUBJECT IN NINTH GRADE OF JUNIOR HIGH SCHOOL

Aferbina Br Barus [1*], Fasya Maharani Umri[2], Jesika Replesia Rajagukguk[3], Nur Hidayah Aini[4], Yuni Apriani Siahaan[5], Tiur Malasari Siregar [6]

[1*,2,3,4,5,6] Universitas Negeri Medan, Medan, Indonesia
** Corresponding author. Jl. W. Iskandar Psr V Medan Esatate, 20221, Medan, Indonesia*

E-mail: aferbina.4233111007@mhs.unimed.ac.id[1*]
fasyaumri.4233311017@mhs.unimed.ac.id[2]
jesikargg.4232411007@mhs.unimed.ac.id[3]
nurhidayah.4231111025@mhs.unimed.ac.id[4]
yunisiahaan.4233111073@mhs.unimed.ac.id[5]
tiurmalasarisiregar@unimed.ac.id [6]

Received November 17, 2025; Received in revised form February 20, 2026; Accepted March 16, 2026

ABSTRACT
This study aims to examine the quality of daily test items in mathematics on the topic of Two-Variable Linear Equation Systems (SPLDV) for grade IX in the 2024/2025 academic year. The study used a quantitative descriptive approach with data in the form of 16 student answer sheets and question scripts used in daily tests. Data collection was conducted through interviews with mathematics teachers to obtain information regarding the objectives and techniques of question preparation. In addition, documentation techniques were also used by collecting copies of essay question sheets and student answer sheets for analysis. The quality of the question items was analyzed through several aspects, namely validity, reliability, difficulty level, and discriminating power. The results of the analysis showed that of the four questions studied, three met the criteria for validity, while one item was declared invalid. The reliability test obtained a Cronbach's Alpha value of 0.492, proving that the level of consistency between the items was still relatively low. Analysis of the level of difficulty showed that three questions were classified as easy questions with an index between 0.71 and 1.00, while the other item was in the medium category. Meanwhile, the results of the discriminatory power analysis stated that all the questions had good ability to differentiate students' ability levels.
Keywords: item analysis; distinguishing power; reliability; difficulty level; validity

ABSTRAK
Penelitian ini bertujuan untuk mengkaji kualitas butir soal ulangan harian pada mata pelajaran matematika pada materi Sistem Persamaan Linear Dua Variabel (SPLDV) kelas IX tahun ajaran 2024/2025. Penelitian menggunakan pendekatan deskriptif kuantitatif dengan data berupa 16 lembar jawaban siswa serta naskah soal yang digunakan pada ulangan harian. Pengumpulan data dilakukan melalui wawancara dengan guru matematika untuk memperoleh informasi mengenai tujuan dan teknik penyusunan soal. Selain itu, teknik dokumentasi juga digunakan dengan mengumpulkan salinan lembar soal uraian dan lembar jawaban siswa sebagai bahan analisis. Kualitas butir soal dianalisis melalui beberapa aspek, yaitu validitas, reliabilitas, tingkat kesukaran, dan daya pembeda. Hasil analisis memperlihatkan bahwa dari empat butir soal yang diteliti, tiga butir memenuhi kriteria valid, sedangkan satu butir lainnya dinyatakan tidak valid. Pada uji reliabilitas diperoleh nilai Cronbach's Alpha sebesar 0,492 yang membuktikan bahwa tingkat konsistensi antarbutir soal masih tergolong rendah. Analisis tingkat kesukaran memperlihatkan bahwa tiga butir soal diklasifikasikan sebagai soal mudah dengan indeks antara 0.71 hingga 1.00, sedangkan satu butir lainnya berada pada kategori sedang. Sementara itu, hasil analisis daya

EMTEKA: Jurnal Pendidikan Matematika
Volume 7, No. 1, 2025. 149-161
DOI: https://doi.org/10.24127/emteka.v6i1.7474

e-ISSN 2746-5594
p-ISSN 2746-5608

pembeda menyatakan bahwa seluruh butir soal memiliki kemampuan yang baik dalam membedakan tingkat kemampuan siswa.
***Kata kunci**: analisis butir soal; daya pembeda; reliabilitas; tingkat kesukaran; validitas*

Introduction

Mathematics education plays a very important role in education because it is related to the development of logical, critical, and systematic thinking skills in students. However, in practice, mathematics is often considered a difficult subject and is less popular among some students. Several studies show that students' interest and motivation in learning mathematics are still relatively low, which results in low student engagement in the learning process (Kurniawan & Djidu, 2021). As a result, mathematics is often perceived as a subject that is complicated and intimidating by many students. Nevertheless, mathematics remains a subject that plays a very important role and provides various uses in everyday life. This affirmation is also stated in Law of the Republic of Indonesia Number 20 of 2003 concerning the National Education System, specifically in Article 37, which states that mathematics is a subject taught at every level of education.

In mathematics learning activities in the classroom, evaluation is an important component that is integrated into the overall learning activities. Through evaluation, educators can determine the level of achievement of learning objectives and assess the effectiveness of the learning process that has been implemented. Sugiyono (2015) emphasizes that evaluation is a stage used to assess the extent to which the plans that have been prepared can be implemented and the extent to which the program objectives have been achieved. In line with this opinion, Magdalena dkk., (2020) argue that evaluation is the activity of collecting, analyzing, and interpreting data to determine the level of student achievement of predetermined learning objectives. Thus, evaluation activities are very important in mathematics learning to monitor the learning process and outcomes achieved by students.

One form of assessment that is often used in mathematics learning is daily tests, which are usually given after students have studied a particular subject. Assessment through daily tests is carried out to obtain an overview of students' level of understanding and their ability to master the material that has been studied. In addition, the results of daily tests are also used as feedback for teachers to assess the success of the learning methods used. However, in practice, the implementation of daily tests in schools often focuses only on giving grades without analyzing the quality of the questions used. As a result, the evaluation results obtained may not accurately reflect the students' abilities.

In conducting daily tests as an evaluation activity, appropriate assessment tools or techniques are needed so that the evaluation process can run systematically and purposefully. The assessment techniques applied can take the form of tests or non-tests as a means of obtaining information related to student learning achievement. In this case, according to (Magdalena dkk., 2021) it is

e-ISSN 2746-5594
p-ISSN 2746-5608

necessary to plan regularly and continuously in the stages of preparation, test and non-test question creation, implementation, observation, and assessment. This is so that teachers can evaluate and guide student development in accordance with the curriculum being implemented. According to Zainal Arifin (2016), tests are one of the tools that can be used in the implementation of assessment activities, where tests are divided into a number of questions or a collection of tasks that students must complete so that teachers can assess various aspects of student behavior. The results of these evaluation activities will produce student scores. Thus, it is necessary to analyze the test items with the aim of obtaining an overview of the quality level of each test item used as a tool or assessment technique in evaluation activities.

The quality of questions can be determined through a method that involves examining the quality of each question item. The examination of question items aims to obtain an overview of the quality of each question by reviewing its strengths and weaknesses, so that it can be used as a basis for improving or revising questions that still need to be corrected (Usman dkk., 2022). Through this analysis, teachers can identify various problems in the questions, such as an inappropriate level of difficulty, the accuracy of the answer keys, and the ability of the questions to differentiate between the varying levels of ability of the students. Therefore, item analysis can help teachers in developing higher quality evaluation instruments for the next learning process.

The quality of evaluation instruments, especially questions used in daily tests, greatly determines the accuracy in measuring students' abilities in mathematics learning. Therefore, the questions used need to undergo an analysis process to ensure that the instruments are truly capable of assessing the expected competencies. Question quality analysis is generally carried out by considering a number of indicators, namely validity, reliability, discriminating power, and level of difficulty (Tarmizi dkk., 2020; Prameswari & Pradani, 2021). Validity describes the ability of a question to measure the competencies that should be the target of measurement, while reliability relates to the stability or consistency of the measurement results obtained. Discrimination describes the ability of a question to distinguish between students' ability levels, while difficulty level describes the difficulty category of a question, namely easy, medium, or difficult. Therefore, analysis of these four aspects needs to be carried out so that the evaluation instruments used are of good quality and able to provide a more accurate picture of students' abilities.

Based on research conducted by Masullah, dkk. (2024), an analysis of the final semester mathematics exam questions for seventh grade at SMP Negeri 6 Praya Timur showed that the quality of the test instruments used still needs to be improved. The results of the study showed that only a small portion of the questions met the validity criteria, while most of the others did not meet the requirements for valid questions. In addition, the reliability of the test showed a low level of consistency. In terms of difficulty, most of the questions were in the easy category, some were in the medium category, and only a few were classified as difficult. In terms of discriminating power, the majority of the questions also showed a low ability to distinguish between students with different levels of ability. These findings provide an important picture of the quality of evaluation

e-ISSN 2746-5594
p-ISSN 2746-5608

instruments in mathematics learning. However, most previous studies have focused more on questions used in final assessments, such as semester exams or school exams. Therefore, this study was conducted to complement existing studies by analyzing the quality of questions in more routine assessments, namely daily tests, by reviewing aspects of validity, reliability, level of difficulty, and discriminating power so as to provide an additional picture of the quality of evaluation instruments used in mathematics learning.

However, in practice at schools, daily tests used as learning evaluation tools often do not undergo a systematic analysis of question quality. This condition has the potential to cause the evaluation results obtained to not fully represent the students' actual abilities. Based on these conditions, it is necessary to examine the questions used in daily tests to determine their quality. In line with this, this study was conducted to assess the quality of ninth-grade mathematics daily test items by considering four main indicators, namely validity, reliability, difficulty level, and discriminating power.

Research Methods

The method used in this study was descriptive research conducted using a quantitative approach. According to (M. Sari dkk., 2023), quantitative descriptive research uses secondary data collection, which is collecting information or data indirectly by observing the object in question. Then, after collecting a number of supporting journals relevant to the topic of discussion, an analysis of the journal content was carried out. This research was conducted at SMP Negeri 3 Berastagi Class IX in the 2024/2025 academic year with the aim of understanding the quality of daily test questions for mathematics subject SPLDV material. In this study, the samples analyzed included 16 student daily test answer sheets and test question sheets. The procedures for collecting information in this study were interviews and documentation. The researcher interviewed mathematics teachers, which yielded information about the objectives, material, and question design techniques used by the teachers. The researcher collected data through documentation in the form of written materials, namely the essay questions and students' daily test answer sheets. After collecting the data, the researcher assessed the students' daily test answer sheets and analyzed the quality of each essay question using Microsoft Excel software and formulas. The research techniques used in this study were:

Item Validity Analysis

An item is considered to have a high level of validity if it is able to accurately perform its measurement function or produce data that is in line with the predetermined measurement objectives, which means that the measurement values accurately capture the facts or real-world conditions of the object being measured. For an assessment tool to be considered valid, it must be accurate with the concept being evaluated in order to evaluate what should be evaluated. A test may not be suitable for different purposes or decisions even though it is valid for one purpose (Ramadhan dkk., 2024). The product moment correlation technique is used to perform validity analysis. According to (R. P. Sari, 2022), validity testing through the Pearson Product Moment correlation test can be done by applying the formula:

EMTEKA: Jurnal Pendidikan Matematika
Volume 7, No. 1, 2025. 149-161
DOI: https://doi.org/10.24127/emteka.v6i1.7474
e-ISSN 2746-5594
p-ISSN 2746-5608

$$r_{xy} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][(n\sum y^2) - (n\sum y)^2]}} \quad (1)$$

Explanation:
r_{xy} = correlation coefficient of variable x with variable y
n = number of respondents
$\sum xy$ = total between the multiplication of item scores and total scores
$\sum x$ = total item scores
$\sum y$ = total overall scores
$\sum x^2$ = sum of squares of item scores
$\sum y^2$ = sum of squares of total scores

The r_{xy} value is the correlation coefficient for each item. This r_{xy} in value is then compared with the product moment r_{table} coefficient at a significance level of 0.05. An item is considered valid if it meets the criterion $r_{count} > r_{table}$.

Item Reliability Analysis

Reliability in research is a measure of the consistency of an instrument in repeatedly measuring the same construct (Sugiono dkk., 2020). In this study, reliability testing was conducted using one method, namely Cronbach's Alpha coefficient, which is used to measure the internal consistency of the research instrument. Cronbach's Alpha coefficient was used to determine the extent to which each item was related to one another in measuring the same variable. Cronbach's Alpha values range from 0 to 1. The closer the value is to 1, the higher the internal consistency of the instrument. An instrument is considered reliable if the Cronbach's Alpha result is 0.6 or higher, while a value below 0.6 indicates that the instrument is not reliable.

The Cronbach's Alpha formula used is:

$$r_{11} = \left(\frac{n}{(n-1)}\right)\left(1 - \frac{\sum \sigma_b^2}{\sigma_t^2}\right) \quad (2)$$

Explanation:
r_{11} = reliability value
n = number of questions
$\sum \sigma_b^2$ = total variance score for each question
σ_t^2 = total variance score for all questions

With this value, it is hoped that the instrument used will have a good level of consistency in measuring the research variable.

Analysis of Item Difficulty Level

According to (Fatimah & Alfath, 2019) in determining the difficulty level of an item, the thing to consider is the opportunity for students to give correct and appropriate answers at a specific ability level. This is also useful for identifying whether the item is easy, medium, or difficult. In their research, Fatimah and Alfath classified questions based on a difficulty index represented by P, which stands for Proportion, with values ranging from 0.00 to 1.00. The results of the difficulty

e-ISSN 2746-5594
p-ISSN 2746-5608

index calculation are then grouped into three types, namely easy, medium, and difficult. According to (Hamimi dkk., 2020), the range of question difficulty is written in Table 1.

Table1. Question Difficulty Index Criteria

Range	Category
0.00 – 0.30	Difficult
0.31 – 0.70	Moderate
0.71 – 1.00	Easy

Based on Table 1, the difficulty index is used to determine the level of difficulty of a question given to students. The difficulty index value describes the level of ease or difficulty of a question for students. A low index value indicates that the item is relatively difficult to complete, while a higher value indicates that the item tends to be easy for students to complete. Based on these values, items can be classified into three difficulty categories: difficult, moderate, and easy. This classification is done to assess the balance of difficulty levels of the items used in the study so that the instruments used can provide an accurate picture..

According to Antony in (Hamimi dkk., 2020), the difficulty level of essay questions can be determined using the following formula:

$$mean = \frac{total\ respondent\ scores\ on\ the\ question}{number\ of\ respondents} \qquad (3)$$

Furthermore, the difficulty level is obtained using the following formula:

$$Difficulty\ Level = \frac{mean}{maximum\ score\ set} \qquad (4)$$

Terd There are two main functions of the difficulty level of test items, namely for teaching evaluation and for teachers. For teaching evaluation, the difficulty level helps in identifying concepts that need to be re-taught, evaluating the strengths and weaknesses of the curriculum, providing feedback to students, and ensuring that the tests created are accurate and relevant. For teachers, this analysis is useful for identifying concepts that need to be repeated in learning and providing students with information about their learning achievements. It also allows teachers to see the focus of the curriculum and detect any questions that may be biased.

Item Discrimination Analysis

According to (Ndiung & Jediut, 2020), the indicator of item discrimination for essay questions (Polytomous Test) is measured using the Ferguson formula, as follows:

$$d = \frac{N^2 - \sum {f_i}^2}{N^2 - \frac{N^2}{n+1}} = \frac{(n+1)(N^2 - \sum {f_i}^2)}{nN^2} \qquad (5)$$

Explanation:

EMTEKA: Jurnal Pendidikan Matematika
Volume 7, No. 1, 2025. 149-161
DOI: https://doi.org/10.24127/emteka.v6i1.7474
e-ISSN 2746-5594
p-ISSN 2746-5608

d = discrimination index
N = number of respondents
f_1 = frequency of each test score
n = number of items

Based on the value of the discrimination index coefficient, it is divided into four categories, namely:

Table2. Discrimination Index Criteria

Coefficient	Category
0.40 – 1.00	Good
0.30 – 0.39	Moderate (no revision needed)
0.20 – 0.29	Needs revision
-1.00 – 0.19	Poor

Based on Table 2, the discrimination coefficient is used to indicate the ability of an item to identify differences between students with high abilities and students with low abilities. A higher discrimination coefficient value indicates that the item is more effective in distinguishing between students' abilities. Conversely, if the discrimination coefficient value is low or even negative, then the item is less effective in showing differences in student abilities and therefore needs to be revised or not used again. Therefore, in order to determine whether the items used in the study have met the criteria for good quality, a discrimination analysis needs to be carried out.

Results and Discussion

Item Validity Analysis

This study uses data from daily mathematics tests on the subject of Two-Variable Linear Equation Systems (SPLDV) for ninth grade students at SMP Negeri 3 Berastagi in the 2024/2025 academic year. The data consists of essay questions and answer sheets from 16 students. Details of the scores obtained by the students can be seen in Table 3.

Table3. Respondents' Daily Test Score Data

Respondents	Score for Each Question				Total
	Question 1	Question 2	Question 3	Question 4	
1	4	2	2	4	12
2	5	4	5	4	18
3	5	5	5	3	18
4	5	4	5	3	17
5	4	4	5	4	17
6	4	5	5	3	17
7	5	5	5	3	18
8	3	4	5	3	15
9	4	5	5	3	17
10	3	4	5	3	15
11	5	3	5	3	16
12	5	5	5	3	18

e-ISSN 2746-5594
p-ISSN 2746-5608

13	5	5	5	5	20
14	5	5	5	5	20
15	5	5	5	3	18
16	4	5	4	2	15

Decisions are based on the value of r_{count} (*Corrected Item-Total Correlation*) $\geq r_{table}$ of 0.497, for $df = 16 - 2 = 14$ and $\alpha = 0.05$. The following are the results of the item validity analysis with a total of 4 essay questions.

Table4. Results of Item Validity Analysis

Question	r_{count}	r_{table}	Criteria
1	0.656	0.497	Valid
2	0.724	0.497	Valid
3	0.715	0.497	Valid
4	0.426	0.497	Not valid

Referring to Table 4, from the validity analysis conducted, there are three items that meet the criteria of $r_{count} > r_{table}$, namely item number 1 with $r_{count} = 0.656$, item number 2 with $r_{count} = 0.724$, and item number 3 with $r_{count} = 0.715$, which means that these three items are valid. Meanwhile, item number 4 does not meet the criteria of $r_{count} > r_{table}$ dengan $r_{count} = 0.426$, which means that this item is considered invalid.

Item Reliability Analysis

Based on Table 3. Data on Respondents' Daily Test Scores, the total variance of all respondents' scores is 4.0625. This total variance indicates the degree of variation between respondents' total scores in answering all questions. The greater the total variance value, the more diverse or different the total scores between respondents are, which indicates variation in understanding or perception of the material. Each question has its own variance value that is calculated to see the extent to which each question can differentiate between respondents. The variance value for each question is 0.529167 for the first question, 0.783333 for the second question, 0.6 for the third question, and 0.65 for the fourth question. The variance for each question is then added up to 2.5625, which is the total variance value for the questions.

According to (Zulpan & Rusli, 2020) reliability is the process of measuring the trustworthiness of results. Therefore, the amount of item variance in these questions plays an important role in calculating the reliability of the instrument because it provides an overview of the extent of differences or variations in each question in separating or distinguishing respondents based on their abilities or understanding. If the item variance is too low, this may indicate that the questions in the instrument are not yet effective enough in distinguishing respondents. Conversely, a relatively high item variance value indicates that the questions are able to detect differences among respondents, which is an initial indication that the instrument has variation in responses.

Furthermore, in testing the reliability of the instrument, Cronbach's Alpha coefficient was used as a measure of internal consistency between items. Cronbach's Alpha is one of the techniques often applied in quantitative research to determine whether the instrument is consistent in measuring the same construct.

e-ISSN 2746-5594
p-ISSN 2746-5608

This internal consistency is very important because a reliable instrument will produce stable results when used again in similar situations.

Cronbach's Alpha is calculated by entering the item variance and total variance values into the formula. Based on calculations using the formula, the reliability value obtained is 0.492308. This result is still below the minimum threshold generally set in research, which is 0.6. If the reliability value is above 0.6, it indicates that the questions have a sufficiently strong relationship or correlation, while a reliability value below 0.6 indicates that the consistency between the questions is still low. Therefore, the instrument is deemed unreliable.

In this study, the reliability value obtained was 0.492308. This figure indicates that the instrument used has not yet reached the established reliability standard, so that the items are still inconsistent in measuring the same construct and are not yet able to produce stable measurement results between respondents. The low reliability value can be attributed to a number of factors, including the poor quality of the questions, the presence of questions that may not be relevant to the construct being measured, or differences in respondents' interpretations of the questions presented. Therefore, further evaluation of this instrument is needed to improve its reliability. This evaluation may include revising the questions, rearranging them, or even removing questions that are not related to the main construct being measured. Overall, the low reliability score indicates that this instrument needs further improvement to produce more accurate and consistent data.

Analysis of Item Difficulty Level

When analyzing the difficulty of each question item, the mean value is needed in calculating the level of difficulty. The data is presented in Table 5.

Table5. Analysis of Item Difficulty Level Based on Respondent Scores

Respondent	Question 1	Question 2	Question 3	Question 4
1	4	2	2	4
2	5	4	5	4
3	5	5	5	3
4	5	4	5	3
5	4	4	5	4
6	4	5	5	3
7	5	5	5	3
8	3	4	5	3
9	4	5	5	3
10	3	4	5	3
11	5	3	5	3
12	5	5	5	3
13	5	5	5	5
14	5	5	5	5
15	5	5	5	3
16	4	5	4	2
Total	71	70	76	54
Mean	4.4375	4.4375	4.75	3.375
Difficulty Level	0.8875	0.8875	0.95	0.675

e-ISSN 2746-5594
p-ISSN 2746-5608

Category	Easy	Easy	Easy	Moderate

Referring to Table 5, difficulty index analysis was used to assess the level of difficulty of each item given to respondents. The difficulty index was calculated by comparing the average score of respondents on each item with the maximum score that could be achieved. If the difficulty index value increases, then the level of ease of the question for respondents is also higher, while a lower value indicates that the question is more difficult. Through this analysis, the questions can then be divided into easy, medium, and difficult levels of difficulty, so that the appropriateness of the difficulty level of the questions in accurately measuring student ability can be determined.

To determine the level of difficulty of each question, a difficulty index calculation is used, which is obtained from a comparison between the average student score and the maximum score that can be achieved. The index value is then used as the basis for determining the difficulty category of the question. Questions with an index value above 0.70 are classified as easy, values between 0.31 and 0.70 berada pada kategori sedang, sedangkan nilai di bawah 0.30 are classified as moderate, while values below 0.30 indicate that the question is difficult. Details of the difficulty level of each question are listed in detail in Table 6.

Table 6. Results of Item Difficulty Analysis

No	Item Category	Item Number	Total	Percentage
1	0.31 – 0.70 Moderate	4	1	25%
2	0.71 – 1.00 Easy	1, 2, 3	3	75%

Referring to Table 6, of the four questions analyzed, three questions (75%) were in the easy category. According to (Arikunto dalam Ningrum dkk., 2023), a question is considered good if its difficulty level is in the range of 0.31–0.70. The research findings show that the quality of the questions used does not fully meet the expected criteria. This can be seen from the number of questions classified as easy with a difficulty index value in the range of 0.71 to 1.00. Of the four questions analyzed, only one question was in the medium category, namely question number 4 related to the material on Two Variable Linear Equation Systems (SPLDV). This condition indicates that some students still have difficulty in answering these questions, so students' understanding of SPLDV material needs to be improved.

Sudijono (Fitriani, 2021) discusses several possible actions that can be taken after analyzing the difficulty level of test items, including:

a. Items with a good (moderate) difficulty level should be stored in a test item bank for future use.
b. Questions in the difficult category have three possible follow-ups:
    1. Deleted and not used again in the next test.
    2. Reviewed to find the cause of students' difficulties in answering, such as ambiguous sentences or inaccurate data. After being revised, the questions can be archived and stored in the question bank for future tests.

e-ISSN 2746-5594
p-ISSN 2746-5608

3. They can be saved for use in highly selective tests, where not all students are expected to pass.

c. Easy questions also have three possible follow-ups:
   1. They can be deleted and not used again in subsequent tests.
   2. Reviewed to identify aspects that make the question too easy, such as answer options that are easy to guess. The question can be revised and stored again in the question bank.
   3. Reused in tests with more lenient passing standards, where most students are expected to pass.

Item Discrimination Analysis

Analysis of discriminating power shows that the items used in the research instrument are able to distinguish students' abilities well. The complete results of the test are shown in Table 7.

Table7. Results of Item Discrimination Analysis

Question	Discrimination Index	Explanation
1	0.642	Good
2	0.896	Good
3	0.558	Good
4	0.890	Good

Referring to Table 7, discriminating power testing was conducted to assess the ability of each question to show differences in student abilities. The discriminating power index value serves as an indicator of the quality of questions in distinguishing student ability levels, where a higher value indicates better question ability. Based on the available calculations, all questions showed a discrimination index value in the good category. Thus, these questions were considered effective in identifying differences in student abilities and suitable for use as research instruments.

The results of the analysis in this study indicate that most of the items met the validity criteria with a percentage of 75%, although the reliability of the instrument was still relatively low. In terms of difficulty level, most of the items were in the easy category and some were in the moderate category, while in terms of discriminating power, all items were classified as good, thus being able to distinguish students' abilities. These results are related to the study by Masullah, dkk. (2024) which also analyzed the quality of mathematics items based on indicators of validity, reliability, difficulty level, and discriminating power. However, that study showed that most items still did not meet the validity criteria and the majority had weak discriminating power. Compared to these findings, this study shows that the analyzed items have a higher level of validity and better discriminating power, thus providing an indication that the quality of the items in the daily tests analyzed in this study is relatively more capable of distinguishing students' abilities.

Conclusion and Suggestion

e-ISSN 2746-5594
p-ISSN 2746-5608

Based on the analysis of the ninth grade mathematics daily test items, it can be concluded that the quality of the test instruments used still needs improvement in several aspects. In terms of validity, most of the test items have met the criteria with a percentage of 75%, while the other 25% have not met the criteria. In terms of reliability, Cronbach's Alpha value of 0.49 shows that the test instrument as a whole does not yet have an adequate level of consistency. In terms of difficulty, 75% of the questions are in the easy category, while the remaining 25% are in the medium category. In terms of discriminating power, all questions (100%) are classified as good so that they can distinguish between students' ability levels. In general, although most of the items have met the validity criteria and have good discriminating power, the test instrument still needs to be improved so that its reliability can be increased and its quality as a learning evaluation tool can be optimized.

In line with the conclusions drawn by the author, it is hoped that teachers can develop questions of better quality if the questions are adequate, and improve questions that are still of poor quality. The author advises students to increase their activity in studying the material by utilizing various learning resources available on various learning platforms. In addition, students are also expected to be more courageous in asking questions to teachers when there is material that they have limited mastery of during the learning process. In its implementation, this study still faces several limitations. The number of respondents involved in the study was not too large, and the items analyzed were also limited because they consisted of only four essay questions. Therefore, future studies are recommended to involve more participants and use a more diverse variety of questions so that the analysis of the quality of the instruments can provide a sufficiently broad and in-depth picture.

Reference

Fatimah, L. U., & Alfath, K. (2019). Analisis Kesukaran Question, Daya Pembeda dan Fungsi Distraktor. *Jurnal Komunikasi dan Pendidikan Islam*, *8*(2), 37–64.

Fitriani, N. (2021). Analisis Tingkat Kesukaran, Daya Pembeda, dan Efektivitas Pengecoh Question Pelatihan Kewaspadaan Kegawatdaruratan Maternal Dan Neonatal. *Paedagoria : Jurnal Kajian, Penelitian dan Pengembangan Kependidikan*, *12*(2), 199–205. https://doi.org/10.31764/paedagoria.v12i2.4956

Hamimi, L., Zamharirah, R., & Rusydy. (2020). Analisis Butir Question Ujian Matematika Kelas VII Semester Ganjil Tahun Pelajaran 2017/2018. *MATHEMA JOURNAL*, *2*(1), 57–66. https://doi.org/10.33365/jm.v2i1.459

Kurniawan, R., & Djidu, H. (2021). Kemampuan Literasi Matematis Siswa: Sebuah Studi Literatur. *EDUMATIC: Jurnal Pendidikan Matematika*, *2*(1), 24–30. https://doi.org/10.21137/edumatic.v2i01.468

Magdalena, I., Fauzi, H. N., & Putro, R. (2020). Pentingnya Evaluasi Dalam Pembelajaran Dan Akibat Memanipulasinya. *Bintang : Jurnal Pendidikan dan Sains*, *2*(2), 244–257. https://doi.org/10.30640/dewantara.v2i1.722

Magdalena, I., Fauziah, S. N., Faziah, S. N., & Nupus, F. S. (2021). Analisis Validitas, Reliabilitas, Tingkat Kesulitan Dan Daya Beda Butir Question Ujian Akhir Semester Tema 7 Kelas III SDN Karet 1 Sepatan. *BINTANG : Jurnal Pendidikan*

e-ISSN 2746-5594
p-ISSN 2746-5608

dan Sains, *3*(2), 198–214. https://ejournal.stitpn.ac.id/index.php/bintang

Masullah, B. D., Zuhry, L. H., Usman, L. H., & Maulana, L. M. G. (2024). Analisis Butir Question Ujian Akhir Semester Mata Pelajaran Matematika Smp Negeri 6 Praya Timur. *ELIPS: Jurnal Pendidikan Matematika*, *5*(2), 152–161. http://journal.unpacti.ac.id/index.php/ELIPS

Ndiung, S., & Jediut, M. (2020). Pengembangan Instrumen Tes Hasil Belajar Matematika Peserta Didik Sekolah Dasar Berorientasi Pada Berpikir Tingkat Tinggi. *Premiere Educandum : Jurnal Pendidikan Dasar dan Pembelajaran*, *10*(1), 94–111. https://doi.org/10.25273/pe.v10i1.6274

Ningrum, W. A., Rahmawati, R. C., Minarti, I. B., & Mekar, R. M. (2023). Analisis Butir Question Ulangan Harian Pembelajaran IPA Pada Kelas VIII Di SMPN 21 Semarang. *Educatio: Jurnal Ilmu Kependidikan*, *18*(1), 91–101. https://doi.org/10.29408/edc.v18i1.18291

Prameswari, A. F., & Pradani, R. A. (2021). Analisis Question Ulangan Harian Mata Pelajaran Bahasa Indonesia Siswa Kelas X SMA N 1 Jetis Bantul. *Lingua Rima: Jurnal Pendidikan Bahasa dan Sastra Indonesia*, *10*(1), 79–84. https://doi.org/10.31000/lgrm.v10i1.4092

Ramadhan, M. F., Siroj, R. A., & Afgani, M. W. (2024). Validitas and Reliabilitas. *Journal on Education*, *6*(2), 10967–10975. https://doi.org/10.31004/joe.v6i2.4885

Sari, M., Rachman, H., Astuti, N. J., Afgani, M. W., & Siroj, R. A. (2023). Explanatory Survey dalam Metode Penelitian Deskriptif Kuantitatif. *Jurnal Pendidikan Sains dan Komputer*, *3*(1), 10–16. https://doi.org/10.47709/jpsk.v3i01.1953

Sari, R. P. (2022). Analisis Butir Question Ujian Akhir Semester Genap Mata Pelajaran Matematika Kelas XI SMK Negeri 1 Percut Sei Tuan T.A. 2019/2020. *Jurnal Penelitian Pendidikan MIPA (JP2MIPA)*, *07*(01), 91–99. https://doi.org/10.32696/jp2mipa.v7i1.1360

Sugiono, Noerdjanah, & Wahyu, A. (2020). Uji Validitas dan Reliabilitas Alat Ukur SG Posture Evaluation. *Jurnal Keterapian Fisik*, *5*(1), 55–61. https://doi.org/10.37341/jkf.v5i1.167

Tarmizi, P., Setiono, P., Amaliyah, Y., & Agrian, A. (2020). Analisis Butir Question Pilihan Ganda Tema Sehat Itu Penting Kelas V SD Negeri 04 Kota Bengkulu. *ELSE (Elementary School Education Journal)*, *4*(2), 124–132. https://doi.org/10.30651/else.v4i2.7090

Usman, Wiwid, & Sulisti, H. (2022). Analisis Butir Question Ulangan Matematika Semester Genap Kelas XI SMA Negeri 2 Pulau Maya. *Jurnal Al 'Adad: Jurnal Tadris Matematika*, *1*(2), 34–43. https://doi.org/10.24260/add.v1i2.1133

Zulpan, Z., & Rusli, A. (2020). Validitas Dan Reliabilitas Instrumen Penilaian Membaca Short Functional Text Pada Siswa SMP Kelas VIII. *Jurnal Pendidikan Guru*, *1*(1), 86–95. https://doi.org/10.47783/jurpendigu.v1i1.66