Quality Analysis of Mathematical Representation Test Instruments in Junior High School Statistics

Nyoman Gita Gayatri[1], Fadhlur Rohman[2], Angga Syaputra[3], Agung Putra Wijaya[4*)], Chika Rahayu[5]
[1,2,3,4,5] Lampung University
*) agung.wijaya@fkip.unila.ac.id

Abstract

This study aims to analyze the quality of an essay test instrument measuring students' mathematical representation ability on eighth-grade statistics material, which was administered to ninth-grade students at SMP Negeri 1 Bandar Lampung. The analysis covers item validity, reliability, difficulty level, and discrimination index. This study employed a quantitative approach with a descriptive method. The research subjects were 27 ninth-grade students, while the object of the study consisted of six essay items on statistics that were developed based on indicators of mathematical representation ability. The results show that only three items (numbers 2, 4, and 6) meet the validity criteria, while items 1, 3, and 5 are invalid. The reliability coefficient of the instrument is 0.685, which falls into the moderate category and is therefore acceptable. All items are categorized as easy based on the difficulty index analysis. However, the discrimination index analysis indicates that all items have low discrimination power and thus fail to distinguish between high- and low-ability students. These findings imply that the instrument requires comprehensive revision, particularly in terms of difficulty level and item quality, in order to more accurately measure students' mathematical representation ability in statistics.

Keywords: Mathematical representation ability; Essay test; Item analysis; Instrument quality; Statistics assessment

Introduction

Evaluation is a fundamental component of the learning process that determines the extent to which learning objectives have been achieved. Through evaluation, educators collect data and information about students' learning abilities to assess program implementation and ensure that educational objectives are realized (Phafiandita et al., 2022). In mathematics education, evaluation holds a particularly strategic position given that mathematics demands conceptual accuracy, logical reasoning, and problem-solving competence. Therefore, assessment instruments must be of high quality to provide an accurate picture of students' mathematical competence (Maulana, 2022). Quality instruments must satisfy several criteria: validity, reliability, appropriate difficulty level, and adequate discrimination power (Mustafa & Masgumelar, 2022). Building upon the need for valid and reliable assessment instruments in mathematics, statistics emerges as a fundamental topic that requires careful evaluation to accurately measure students conceptual understanding.

Statistics is a core sub-topic in Grade 8 mathematics that encompasses data collection, presentation, analysis, and interpretation. Statistics is crucial for developing students' critical thinking skills, enabling them to draw evidence-based conclusions and make informed decisions. To effectively evaluate student mastery of statistics concepts, educators must use assessment instruments that accurately reflect students' conceptual understanding. While developing questions aligned with competency indicators is essential, empirical analysis of instruments is equally necessary to ensure that assessment items are valid and reliable.

However, findings from previous research conducted by Dewi et al (2019) reveal that many assessment instruments used in schools do not fully meet quality criteria. Some items are too easy or too difficult, while others fail to discriminate between students of varying ability levels. Additionally, many items are statistically invalid or unreliable, resulting in inaccurate assessment information that can hinder appropriate educational decision-making (Maulana, 2022). Therefore, systematic instrument analysis is essential to improve assessment quality.

In fact, previous studies Dewi et al (2019) have found that many assessment instruments used in schools do not fully meet the criteria for good test items. Some questions are considered too easy, too difficult, or unable to distinguish between high-ability and low-ability students. In addition, many questions are statistically invalid or unreliable. These problems have the potential to produce inaccurate assessment information and hinder the process of making appropriate decisions regarding student development. Therefore, instrument analysis is an important step in improving the quality of assessment.

The use of essay format was intentional, as it enables students to demonstrate their mathematical thinking, problem-solving procedures, and conceptual understanding more comprehensively than multiple-choice questions (Ekayanti & Mahmudah, 2024).

The specific research questions guiding this study are:

1. To what extent do the essay items meet validity criteria?
2. Is the instrument reliable for measuring mathematical representation ability?
3. What is the difficulty level of each item, and is it appropriate for the target grade level?
4. What is the discrimination power of each item, and can they effectively differentiate students by ability level?

This study has significant implications for improving assessment practices in mathematics education. By systematically analyzing instrument quality, educators can identify weaknesses and develop more effective assessment tools. The findings provide guidance for teachers, schools, and curriculum developers in creating valid, reliable, and appropriately challenging assessment instruments.

More broadly, this study has significant importance for the development of assessment instruments in schools. The findings of this study can be used as a basis for teachers, schools, and educational researchers to improve the quality of assessment instruments, particularly in mathematics. In addition, this study is expected to help schools reflect on and improve the assessment process so that the learning outcomes obtained can reflect students' abilities objectively and accurately. Thus, the analysis of the Mathematics assessment instruments for the statistics sub-material for grade VIII students in class IX.11 at SMP Negeri 1 Bandar Lampung is a strategic step to improve the quality of learning evaluation.

Through this study, it is hoped that a deep understanding of the characteristics of the essay questions used and recommendations for optimizing assessment instruments in mathematics learning at the junior high school level can be obtained. Mathematical representation is a fundamental skill that allows students to organize and communicate their mathematical ideas through visual, symbol, or verbal means. Analyzing instruments specifically designed for this ability is vital, as it ensures the assessments can accurately capture how students bridge the gap between abstract concepts and concrete problem solving. A high quality instrument not only measures correctness but also reveals the dept of a students conceptual understanding and cognitive flexibility.

Method

This study employed a quantitative approach with a descriptive method to analyze instrument quality (Tresnahadi et al., 2022). The research was conducted in the first semester of the 2025/2026 academic year at SMP Negeri 1 Bandar Lampung. This study analyzed an essay test instrument designed for Grade 8 statistics but administered to Grade 9 students at SMP Negeri 1 Bandar Lampung. This approach allowed examination of whether the instrument could reliably measure students' mathematical representation ability at a higher grade level, while also assessing the instrument's suitability for students' cognitive development. The participants were 27 ninth-grade students from class IX.11.

The object of study was six essay test items designed to measure mathematical representation such as representing data in tables, graphs, diagrams, and symbols); (2) development of essay items requiring students to demonstrate multiple forms of mathematical representation, including data organization, visual presentation, calculation procedures, and interpretation; and (3) content validation by mathematics educators to ensure alignment between items and learning objectives ability in Grade 8 statistics material.

In this study, the instrument used was an essay-type test, which allowed students to demonstrate their thought processes, steps taken to solve problems, and their ability to explain concepts in greater depth (Ekayanti & Mahmudah, 2024). The essay format was chosen because it provides a more comprehensive picture of students' understanding of statistical concepts than multiple-choice questions, which only assess the final answer. The essay format used in this instrument provides additional important information about students' ability to give reasons, present calculation steps, and interpret statistical calculation results. In the context of essay tests, instrument analysis not only looks at the correctness of answers but also assesses the quality of the scoring rubric, the clarity of the questions, and the ability of the questions to reveal the depth of students' understanding.

Instrument analysis was conducted through a series of statistical tests, including item validity tests to determine the accuracy of the questions in measuring the intended abilities, reliability tests to examine the consistency of the measurement results, difficulty level tests to determine the ease or difficulty of the questions, and discrimination power tests to assess the ability of the questions to distinguish between students with high and low abilities (Maulana, 2022). In essay instruments, this analysis also considers the consistency of the scoring rubric and the variation in the quality of students' answers. The results of this analysis are expected to provide a comprehensive overview of the quality of the instruments used and provide constructive feedback for teachers in preparing future evaluation questions.

The essay test instrument was developed through the following procedures: (1) creation of an assessment grid based on Grade 8 basic competencies and learning indicators related to mathematical representation (such as representing data in tables, graphs, diagrams, and symbols); (2) development of essay items requiring students to demonstrate multiple forms of mathematical representation, including data organization,

visual presentation, calculation procedures, and interpretation; and (3) content validation by mathematics educators to ensure alignment between items and learning objectives.

The instrument was accompanied by a detailed scoring rubric to ensure consistent and objective evaluation. Each item was scored based on the accuracy of calculations, appropriateness of representation, clarity of reasoning, and completeness of explanation. The maximum total score was 24 points (4 points per item).

Data were collected by administering the essay test to all 27 participants. Student responses were scored according to the established rubric. Data analysis included four components:

a. Validity testing: Item validity was examined using Pearson Product Moment correlation between individual item scores and total test scores at a significance level of $\alpha=0.05$ $\alpha=0.05$. An item was considered valid if the calculated correlation coefficient exceeded the critical value of 0.381 (*with df=25 df=25*) (Dewi et al., 2019).
b. Reliability testing: Instrument reliability was calculated using Cronbach's Alpha coefficient. Reliability values were interpreted as follows: < 0.60 = *poor; 0.60–0.70* = acceptable; *0.70–0.80* = good; > 0.80 = excellent (Arifin, 2017).
c. Difficulty index analysis: Item difficulty was determined by converting the mean score for each item to an index ranging from 0 to 1, categorized as: 0.00–0.32 (difficult); 0.33–0.66 (moderate); 0.67–1.00 (easy) (Dewi et al., 2019).
d. Discrimination power analysis: The discrimination index for each item was calculated by comparing mean scores of the top 27% of students (high-ability group) with the bottom 27% (low-ability group). Items were classified according to the following criteria: ≥0.40 (excellent); 0.30–0.39 (good, acceptable with revision); 0.20–0.29 (fair, needs improvement); ≤0.19 (poor, should be rejected or revised) (Dewi et al., 2019).

All data analysis was performed using IBM SPSS Statistics version 26. The following is an attachment of questions that will be used in the research :

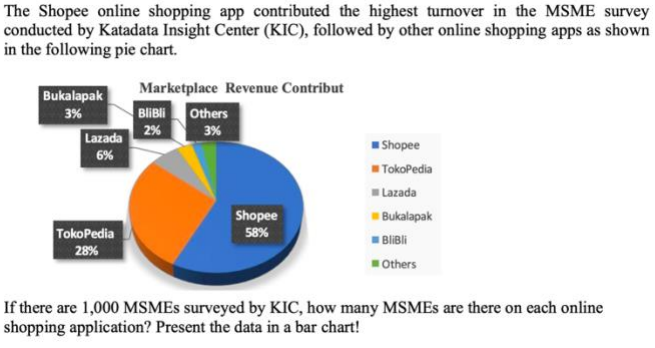

Figure 1. Question number 1

In the data, online buying and selling transactions in the range of 2017 to 2021 have increased significantly. As presented in the following diagram. In what year did the number of online buying and selling transactions increase the highest?

Source: TEMPO

Figure 2. Question number 2

In some types of sports, height is one of the advantages in a game or competition, one of which is football. The height of the Indonesian national team players (in cm) is as follows:

187	187	187	190	178	172	181	196	186	178
182	170	185	174	184	182	170	185	183	177
178	165	181	187	172	171	179	174	174	173
184	186	169	172	175	185	180	176	180	170

What is the average height of the Indonesian national team players? How many people have a height below average?

Figure 3. Question number 3

Take a look at the following picture!

Explain the mode in the library visitor data above!

Figure 4. Question number 4

A family recorded its expenses during the month of April as follows:

Category	Total
Food	IDR 1,500,000
Transportation	IDR 800,000
Education	IDR 700,000
Electricity and water	IDR 400,000
Entertainment	IDR 300,000

If the family's income is IDR 4,000,000 per month, what is the percentage of expenditure for each category?

Figure 5. Question number 5

Reporting from the geeksforgeeks.org page, here are some of the popular online games to play in the world in 2023.

Game Name	Users
PUBG	100.000.000
Minecraft	95.000.000
Apex Legends	50.000.000
Fortnite	45.000.000
League of Legends	20.000.000
Call of Duty	15.000.000

Explain the above data using a line chart!

Figure 6. Question number 6

Data collection was conducted through essay tests and documentation in the form of question grids, scoring guidelines, and student answer sheets. The research procedure began with the preparation of instruments, including the creation of grids based on basic competencies, the preparation of essay questions based on learning indicators, and content validation by teachers or experts to ensure the suitability of the material, context, and question construction. The essay instruments were also equipped with scoring rubrics to facilitate an objective and focused assessment process. Once the instruments were deemed suitable, the questions were administered to students in class IX.11, and the answer sheets were collected for analysis.

Data analysis was conducted through item validity testing using Pearson's Product Moment correlation between each item score and the total test score. Instrument reliability was calculated using Cronbach's Alpha formula, which is appropriate for essay tests. Difficulty level analysis was conducted by looking at the average scores obtained by students on each item, while discriminating power was calculated by comparing the average scores of the top and bottom groups. For essay questions, the assessment of instrument quality also considered the consistency of scoring based on a pre-determined rubric. The results of this analysis were used to assess the quality of each essay question item, whether it was classified as good, needed revision, or was not suitable for use.

Overall, this study provides empirical information regarding the feasibility of essays as an evaluation tool for statistics material. The study was conducted at SMP Negeri 1 Bandar Lampung in the first semester of the 2025/2026 academic year.

Results and Discussion

Validity Test

Item validity was assessed using the Pearson Product Moment correlation, where the critical value (r-table) was *0.381* at $\alpha=0.05\alpha=0.05$ and $df=25df=25$. An item was considered valid if the calculated correlation coefficient (r-calculated) exceeded the critical value (r-table).

Table 1. Validity Test

Item	*r* Calculated	*r* Table	Status
Item 1	0.239	0.381	Invalid
Item 2	0.471	0.381	Valid
Item 3	-0.141	0.381	Invalid
Item 4	0.717	0.381	Valid
Item 5	-0.023	0.381	Invalid
Item 6	0.822	0.381	Valid

The analysis revealed that only three items (2, 4, and 6) met validity criteria, with correlation coefficients of *0.471, 0.717*, and *0.822* respectively. These items demonstrated adequate correlation between item performance and overall test performance, indicating that they effectively measure the construct of mathematical representation ability. Conversely, items 1, 3, and 5 failed to meet validity standards, with correlation coefficients of *0.239*, *-0.141*, and *-0.023* respectively. The negative correlations for items 3 and 5

suggest that student performance on these items was inversely related to overall performance, indicating that these items are not functioning as intended and require substantial revision or replacement.

Reliability Test

Instrument reliability was calculated using Cronbach's Alpha coefficient, which measures the internal consistency of all test items. This coefficient is particularly appropriate for essay tests where items assess multiple dimensions of a construct.

Table 2. Reliability Test

Reliability Statistics	
Cronbach's Alpha	N of Items
.685	3

Source : SPSS Data Analysis Result

The calculated Cronbach's Alpha coefficient of 0.685 indicates that the instrument possesses acceptable internal consistency. While this value falls in the moderate range (0.60–0.70) rather than the good range (0.70–0.80), it is still acceptable for research purposes. The moderate reliability level suggests that while the items share some common variance related to mathematical representation ability, there is also considerable item-specific variance. This moderate consistency is noteworthy given that only three of the six items were found to be valid; had the analysis been limited to valid items only, the reliability coefficient would likely be higher. The moderate reliability, combined with the fact that validity was problematic for half the items, reinforces the need for comprehensive instrument revision.

Difficulty Index

Difficulty index is the degree of difficulty of a question item expressed in numerical form. (Hera et al., 2023), (Arifin, 2017) also adding difficulty index is the third stage of data testing that will be carried out in item analysis. The difficulty index of an item is the percentage or proportion of students who took the test and answered the item correctly.

Difficulty index categories include difficult, moderate, and easy. The following is a breakdown of the difficulty index categories into three groups:

Table 3. Difficulty Index Criteria

Difficulty Index Range	Difficulty Index Categori
0,00 – 0,32	Hard
0,33 – 0,66	Medium
0,67 – 1,00	Eassy

Source: Dewi et al., (2019)

The results of the SPSS data test assessment or scores obtained on the difficulty level test instrument can be seen in the figure.

Table 4. Difficulty Index Test

Item	Mean Score	Difficulty Index	Category
Item 1	3.37	0.84	Easy
Item 2	3.56	0.89	Easy
Item 3	2.07	0.52	Moderate
Item 4	3.18	0.80	Easy
Item 5	3.96	0.99	Easy
Item 6	3.40	0.85	Easy

Source : SPSS Data Analysis Result

The analysis reveals that five of six items fall in the easy category (difficulty index > 0.67), with item 5 being particularly easy (index = *0.99*). Only item 3 achieved a moderate difficulty level (index = *0.52*). This finding indicates that the majority of participants were able to answer most items correctly, suggesting that the instrument may not be appropriately challenging for ninth-grade students. While the items were originally designed for Grade 8, their administration to Grade 9 students appears to have resulted in artificially lowered difficulty, a phenomenon that likely contributes to the poor discrimination power observed across all items.

Discrimination Index

The discrimination power of items was assessed by calculating the difference in mean performance between the top 27% of students (*high-ability group, n=7n=7*) and the bottom 27% (*low-ability group, n=7n=7*) (Magdalena & Jaolis, 2018). Results were interpreted using the following criteria: *≥0.40* = excellent; *0.30–0.39* = good/acceptable; *0.20–0.29* = fair/needs improvement; ≤0.19 = poor/should be rejected (Dewi et al., 2019).
There is a relationship between discriminating power and question quality, which can be classified as follows:

Table 5. Discrimination Index Criteria

Discrimination Index Range	Criteria
0,40 or more	The item is very good and acceptable
0,30 – 0, 39	The item is quite good and acceptable with improvments
0,20 – 0,29	The item is fair and needs discussion. It usually needs improvement and become a target for improvment
0,19- and below	The item is poor, rejected or discarded and replaced with another item

Source : Dewi et al., (2019)

The results of the SPSS data test for the assessment or scores obtained on the discrimination power test instrument can be seen in Table 6.

Table 6. Discrimination index Test

Item	High Group Mean	Low Group Mean	Discrimination Index	Category
Item 1	3.29	3.57	-0.101	Poor
Item 2	3.57	3.71	-0.015	Poor
Item 3	2.29	2.00	0.072	Poor
Item 4	3.57	3.00	0.202	Poor
Item 5	3.86	4.00	-0.042	Poor
Item 6	3.86	2.86	0.286	Fair

A critical finding emerged: all items exhibited discrimination indices below 0.19, indicating poor discrimination power. Most concerning are items 1, 2, and 5, which showed negative discrimination indices, meaning that lower-ability students actually performed better than higher-ability students on these items. This counterintuitive result suggests fundamental problems with item construction or measurement validity. Item 3 and 4 showed minimal positive discrimination (*0.072* and *0.202* respectively), while only item 6 approached the lower bound of acceptable discrimination (*0.286*).

The poor discrimination across all items is directly attributable to the pattern of item difficulty discussed above. Since most items are very easy (indices > *0.80*), students across all ability levels tend to answer correctly, resulting in homogeneous score distributions with minimal variation. This ceiling effect prevents adequate differentiation

between high- and low-performing students. The negative discrimination indices are particularly problematic, as they suggest that lower-performing students paradoxically score higher on certain items, which may indicate ambiguous question wording, inappropriate answer keys, or misalignment between item content and students' conceptual understanding. The low discrimination power is often caused by an imbalance in difficulty levels, specifically when items are too easy. This aligns with the findings in the Setiyawan & Wijayanti (2020) which states that a high-quality instrument must maintain a balanced distribution of difficulty levels to achieve optimal discrimination power.

Conclusion and Suggestion

This study concludes that the essay test instrument analyzed has not yet met adequate quality standards for measuring students' mathematical representation ability in statistics. Although the instrument demonstrated moderate reliability, several items were invalid, overly easy, and showed poor discrimination power. These weaknesses indicate that substantial revision is necessary, particularly in improving item difficulty, clarity, and alignment with representation indicators. Future instrument development should emphasize balanced cognitive demand and stronger representation components to ensure accurate and meaningful assessment outcomes.

References

Arifin. (2017). Kriteria Instrumen Dalam Suatu Penelitian. *Jurnal Theorems (The Original Research Of Mathematics)*, *2*(1), 28–36.

Dewi, S. S., Hariastuti, R. M., & Utami, A. U. (2019). Analisis Tingkat Kesukaran Dan Daya Pembeda Soal Olimpiade Matematika (Omi) Tingkat Smp Tahun 2018. *Transformasi : Jurnal Pendidikan Matematika Dan Matematika*, *3*(1), 15–26. Https://Doi.Org/10.36526/Tr.V3i1.388

Ekayanti, F., & Mahmudah, I. (2024). Efektivitas Penggunaan Essay Pada Evaluasi Pembelajaran Matematika Kelas Iv Di Min 2 Kota Palangka Raya. *Jurnal Ilmiah Pendidikan Guru Madrasah Ibtidaiyah*, *4*(1).

Hera, A. S., Zulhijrah, Nabila, J. L., & Shaleh. (2023). Analisis Instrumen Assesmen : Validitas, Reliabilitas, Tingkat Kesukaran Dan Daya Beda Butir Soal. *Jurnal Ilmiah Pgsd Fkip Universitas Mandiri*, *9*, 2990–2991.

Magdalena, A., & Jaolis, F. (2018). Analisis Antara E-Service Quality, E-Satisfaction, Dan E-Loyalty Dalam Konteks E-Commerce Blibli. *Jurnal Strategi Pemasaran*, *5*(2), 1–11. Https://Publication.Petra.Ac.Id/Index.Php/Manajemen-Pemasaran/Article/View/7190

Maulana, A. (2022). Analisis Validitas , Reliabilitas , Dan Kelayakan Instrumen Penilaian Rasa Percaya Diri Siswa. *Jurnal Kualitas Pendidikan*, *3*(3), 133–139.

Mustafa, P. S., & Masgumelar, N. K. (2022). Kajian Review: Pengembangan Instrumen Penilaian Sikap, Pengetahuan, Dan Keterampilan Dalam Pendidikan Jasmani Dan Olahraga. *Biormatika: Jurnal Iimiah Fakultas Keguruan Dan Iimu Pendidikan*, *8*(1), 31–49.

Phafiandita, A. N., Permadani, A., Pradani, A. S., & Wahyudi, M. I. (2022). Urgensi Evaluasi Pembelajaran Di Kelas. *Jurnal Inovasi Dan Riset Akademik*, *3*(2), 111–121.

Setiyawan, R. A., & Wijayanti, P. S. (2020). Analisis Kualitas Instrumen Untuk Mengukur Kemampuan Pemecahan Masalah Siswa Selama Pembelajaran Daring Di Masa Pandemi. *Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*, *1*(2), 130–139.

Tresnahadi, D. P. T., Sugilar, & Noviyanti, M. (2022). Kontribusi Adversity Quotient dan Motivasi Belajar Matematika Terhadap Prestasi Belajar Matematika Siswa Smk Negeri Se-Kabupaten Buleleng. *Jurnal Impresi Indonesia*, *1*(10), 1025–1031. https://doi.org/10.36418/jii.v1i10.464