ISSN : 2962-9195 E-ISSN : 2963-8992

Journal of Educational Management Research
Vol. 04 No. 02 (2026) : 931-946
Available online at https://serambi.org/index.php/jemr

Educational Quality Management in Elementary Schools: Measuring the Validity and Reliability of Maharah Qira'ah Test Items

Ahmad Mashadar Hilmi[1], Muhammad Syihabul Ihsan Al Haqiqy[2]*, Aminudin[1] , Rahmah Fadhilah Agustina[3]
[1]Sekolah Tinggi Agama Islam Jarinabi, Indonesia
[2]Universiatas Sunan Gresik, Indonesia
[3]Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia
Email : ahamdmashadar@gmail.com

DOI: https://doi.org/10.61987/jemr.v4i2.534

Keywords:
Test Items, Arabic Language, Reading Skills

**Corresponding Author*

ABSTRACT

This study aims to evaluate the quality of test items by analyzing their validity, reliability, difficulty level, discriminating power, and distractor effectiveness in measuring students' learning outcomes. A descriptive quantitative design was employed, involving sixth-grade students of Tarbiyatus Shibyan Elementary School, Duduksampeyan Gresik, during the 2024/2025 academic year. Data were collected through documentation of students' test results and analyzed using SPSS version 25 and Microsoft Excel. The findings reveal that most items meet the criteria for validity, as indicated by r-values exceeding the r-table or significance levels below 0.05. The reliability coefficient (Cronbach's Alpha) also exceeds 0.6, demonstrating strong internal consistency. The difficulty level analysis shows that 68% of the items fall into the moderate category, with no items classified as difficult. However, the discriminating power is predominantly poor, with 72% of items showing indices below 0.19, and 24 items contain ineffective distractors. These results imply the need for improved item construction to ensure more accurate, fair, and balanced assessments in elementary education

Article History:
Received: March 2025; Revised: April 2025; Accepted: June 2025

Please cite this article in APA style as:
Hilmi, A. M., Al Haqiqy, M. S. I., Aminudin, A., & Agustina, R. F. (2025). Educational Quality Management in Elementary Schools: Measuring the Validity and Reliability of Maharah Qira'ah Test Items. *Journal of Educational Management Research, 4*(2), 931-946.

INTRODUCTION

In contemporary education systems, the quality of learning is increasingly recognized as a national priority because it shapes students' intellectual growth and determines their ability to participate effectively in society (Alam & Mohanty, 2023; Kusnanto et al., 2023). One crucial component of quality learning is assessment, which functions not only as a measurement tool but also as a feedback mechanism that guides teachers in determining subsequent

instructional decisions. When assessment is poorly designed, the entire learning cycle is compromised, leading to inaccurate conclusions about students' abilities and ineffective teaching adjustments (Jamil & Sanusi, 2024; Makiyah, 2024; Manshur et al., 2024). Numerous reports highlight that many schools still fail to implement assessments that meet proper measurement standards, which affects educational accountability and the reliability of learning outcomes. These conditions demonstrate that improving assessment quality is not simply a technical concern but a societal necessity, as valid and reliable assessments ensure fair evaluations, support educational equity, and strengthen long-term learning outcomes. Therefore, studies focusing on the quality of teacher-made tests including their validity, reliability, and item characteristics play an essential role in enhancing educational systems and contributing to broader social development.

Despite the central role of assessment in shaping instructional decisions, many educational institutions continue to face persistent challenges in developing high-quality test instruments. The widespread use of teacher-made tests that lack systematic item analysis often leads to the use of flawed instruments that cannot accurately measure students' competencies (Alam & Mohanty, 2023; Kusnanto et al., 2023). This issue becomes more severe when teachers rely solely on previous test formats or personal intuition instead of empirical analysis, resulting in assessment tools that may be too difficult, too easy, or unable to differentiate between high- and low-performing students. In the broader educational landscape, these weaknesses create substantial learning gaps because decisions regarding remediation, enrichment, or curriculum adjustments are often based on misleading test results. Consequently, students may be misclassified, instructional efforts may be misdirected, and overall learning outcomes may remain stagnant (Alam & Mohanty, 2023; Kusnanto et al., 2023). This problem is critical because assessments that fail to meet basic quality standards such as validity, reliability, item difficulty, discrimination, and distractor effectiveness undermine the integrity of the entire educational process. Therefore, addressing assessment quality is indispensable for ensuring accurate measurement and effective learning development.

In many elementary schools, real-world observations suggest that test construction practices still rely heavily on conventional habits rather than systematic measurement principles (Alam & Mohanty, 2023; Kusnanto et al., 2023; Maripaz C. Abas, 2025). Teachers tend to prepare tests independently without conducting item analysis to verify the quality of their instruments, and this practice is widespread across various subjects (Aini & Wahid, 2024; Makiyah, 2024; Mukarromah et al., 2024). As noted by Halik et al. (2019), assessment is an inseparable component of the learning cycle, yet its implementation in schools

often does not reflect this theoretical importance. Many teachers cite limited time, insufficient training, or lack of access to analysis tools as reasons for not evaluating item quality (Barokah, 2025; Jannah & Rizquha, 2025; Muharromah, 2025; Zakiyah, 2025). As a result, numerous test items used in classrooms fail to meet essential measurement standards, producing assessments that do not accurately capture student learning. In addition, inconsistencies in test difficulty, weak discriminatory power, and ineffective distractors further reduce the usefulness of teacher-made tests. Such conditions not only distort students' actual learning achievements but also prevent teachers from making precise decisions for instructional improvement, revealing a substantial gap between assessment theory and classroom practice (Dewi et al., 2025; Khoiroh, 2025; Maripaz C. Abas, 2025; Yakin, 2025).

Previous research has extensively emphasized the importance of item analysis as a foundation for improving the quality of educational assessments. Studies conducted by Purniasari et al. (2021), Putriani et al. (2020), Hasbullah (2020), Al Haqiqy et al. (2024), and Nur Cahyo et al. (2022) consistently reveal that many test items used in schools suffer from weaknesses in discrimination power and distractor effectiveness (Fajariyah, 2025; Hadi & Masuwd, 2025; Indayati, 2025). These findings indicate that large portions of teacher-made tests cannot differentiate student performance accurately and often include distractors that do not function as intended. Although these studies contribute significantly to understanding existing assessment problems, they primarily focus on subjects such as mathematics, science, or Indonesian language. As a result, the literature lacks comprehensive investigations into test quality in other disciplines, especially Arabic language learning. This gap is critical because language assessments possess distinct characteristics particularly in reading comprehension that cannot be evaluated solely using general item-analysis frameworks. Therefore, further research is needed to expand the scope of item-quality investigations into less-studied subject areas to strengthen educational assessment practices.

Additionally, prior studies commonly analyze only two or three components of test quality, such as validity and reliability, without integrating all five essential parameters: validity, reliability, item difficulty, discrimination index, and distractor effectiveness (Alam & Mohanty, 2023; Kusnanto et al., 2023; Wagner, 2025). This partial approach limits the comprehensiveness of conclusions regarding overall test quality (Kholifatunnisak, 2024; Ma'isyah et al., 2024; Nuriyah et al., 2024). Another limitation is the scarcity of research that employs modern digital tools such as SPSS or Excel to assist teachers in conducting empirical item analysis systematically (Kusumaningsih & Riauwanto, 2021; Rostiyanti et al., 2023). Although these tools are widely

available and increasingly accessible, many research studies do not incorporate them, resulting in analyses that may lack precision or practical applicability. Furthermore, the literature rarely addresses assessment challenges in Arabic language learning at the elementary school level, particularly in maharah qira'ah, which requires specific and nuanced evaluation techniques. This absence highlights a significant research gap that needs to be filled to provide educators with empirical evidence and practical models for improving assessment quality. Given these limitations, there is a pressing need for more holistic and technologically supported research on test quality in Arabic reading instruction.

The present study introduces several novel contributions that address these significant gaps. First, it focuses specifically on Arabic language assessment, particularly on maharah qira'ah in elementary schools, an area that remains underrepresented in existing evaluation research. This focus is essential because reading in Arabic requires unique competencies involving script recognition, vocabulary understanding, and contextual interpretation, making the quality of test items especially important. Second, this study integrates all five critical components of item analysis validity, reliability, difficulty level, discrimination index, and distractor effectiveness thus offering a more comprehensive evaluation framework than many previous studies. Third, the research employs SPSS 25 and Excel to conduct empirical item analysis, demonstrating a practical and replicable model that teachers can adopt to improve their classroom-based assessments. By combining subject-specific focus, methodological completeness, and digital analysis tools, this study contributes a state-of-the-art approach that enhances both theoretical understanding and practical assessment management, providing a uniquely comprehensive perspective within the field of Arabic language education.

Based on these considerations, the central research problem of this study concerns the extent to which Arabic maharah qira'ah test items used in elementary schools meet established standards of assessment quality. The study seeks to determine whether these test items demonstrate adequate validity, reliability, appropriate difficulty levels, strong discrimination power, and effective distractors. The underlying argument is that without empirically tested assessment tools, teachers cannot accurately measure students' reading competencies or make informed instructional decisions. Therefore, this research provides a systematic and evidence-based evaluation that addresses a critical gap in both theory and practice. The findings are expected to support teachers in designing higher-quality assessments, enhance the accuracy of measurement in Arabic reading instruction, and contribute to improving overall learning outcomes. Furthermore, the study asserts that integrating comprehensive item analysis into educational management practices can significantly strengthen

assessment systems, offering both practical and theoretical contributions to the ongoing development of evaluation practices in Arabic language education.

RESEACH METHOD

This study employs a quantitative approach with a descriptive research design. A quantitative approach is characterized by a clear and systematic structure that allows researchers to examine phenomena objectively through numerical data (Sugiyono, 2020). Meanwhile, descriptive research aims to present accurate facts without testing hypotheses or establishing causal relationships, focusing instead on describing the characteristics of a particular phenomenon in detail (Creswell, 2021). This design was chosen because the objective of the study is to describe the validity, reliability, difficulty level, and discrimination power of test items, which are best analyzed through numerical indicators and statistical procedures.

The research was conducted at MI Tarbiyatus Shibyan, Sumengko, Gresik. This location was selected because the school had implemented an Arabic language learning program focusing on reading skills, making it relevant to the purpose of evaluating maharah qira'ah test items. Additionally, the institution provided access to students who had already studied the targeted Arabic reading material, ensuring that the test items could be administered appropriately. The availability of supportive teachers and the school's willingness to facilitate research activities also contributed to the selection of this setting. Data were collected using a documentation technique, specifically through students' test results. The instrument consisted of 25 multiple-choice items designed to measure students' Arabic reading competency. The test was administered to 15 students who had completed learning related to the tested material. The documentation method was selected because it allowed the researcher to obtain direct and objective data regarding students' responses, which are essential for conducting item analysis.

The collected data were analyzed using SPSS version 25 to examine the quality of the test items. The item analysis included calculating validity, reliability, item difficulty, discrimination index, and distractor effectiveness. These components were evaluated to determine whether each test item met acceptable measurement standards. The statistical outputs generated by SPSS provided empirical evidence regarding the performance of each item and helped identify items that required revision or elimination. To ensure the accuracy and credibility of the data, several verification steps were undertaken. First, the test items were constructed based on relevant learning materials studied by the students, ensuring content alignment. Second, the scoring and data entry processes were checked manually to prevent errors before running the analysis

in SPSS. Third, the analysis procedures were conducted using standardized statistical formulas embedded in the software, ensuring consistency and objectivity. These steps strengthened the validity of the findings and ensured that the results accurately reflected the quality of the test items.

RESULT AND DISCUSSION

Validity Test

Validity testing is a tool used to measure the validity or appropriateness of the instrument used by researchers to measure and obtain data from respondents. Validity is measured by the Pearson correlation (Product Moment correlation) between the instrument items and the total number of instruments (Zamzaili & Swita, 2021)..

In this study, the validity of the questions was determined by analyzing multiple-choice Arabic language questions using IBM SPSS 25. The R-table and sig. values were compared as validity test tools, which serve to assess the validity of an instrument used in the exam questions.

The basis for this decision is as follows:

1. compare the calculated r value with the table r value
   a. If the calculated r value is greater than the table r value, then the question is declared valid.
   b. If the calculated r value is less than the table r value, then the question is declared invalid..
2. Look at the significance value (sig).
   a. If the significance value is <0.05, then the question is declared valid.
   b. If the significance value is >0.05, then the question is declared invalid.

Validitas butir soal This analysis uses the Biserial Correlation technique (Khofifah et al., 2020). Interpretation of the Point Biserial correlation index shows the validity value in the criteria in Table 1.

Table 1. Validity Table

Test Validity Category	Correlation Coefficient Value
Very high	0,800 – 1,000
Tall	0,600 – 0,799
Currently	0,400 – 0,599
Low	0,200 – 0,399
Very Low	0,000 – 0,199

Based on table 1 about the validity table, it is explained that to achieve a very high test validity category, the correlation coefficient value ranges from

0.800 to 1.000. While high test validity ranges from 0.600 – 0.799. Medium validity, the correlation coefficient value between 0.400 – 0.599. Meanwhile, the test validity is categorized as low and very low when the correlation coefficient value ranges from 0.200 – 0.399 and 0.000 – 0.199 (Arikunto, 2014).

The following are the results of validity calculations using IBM SPSS Statistics 25.

Table 2. IBM SPSS Statistics 25 Calculation Results

No. Question	R Calculate	R Significance	Table (GIS)	Description	Interpretation
1	0,617	0,396	0,014	Valid	High
2	0,857	0,396	0,000	Valid	Very High
3	0,986	0,396	0,000	Valid	Very High
4	0,986	0,396	0,000	Valid	Very High
5	0,421	0,396	0,118	Valid	Medium
6	0, 986	0,396	0,000	Valid	Very High
7	0,798	0,396	0,000	Valid	High
8	0,320	0,396	0,245	Invalid	Low
9	0,986	0,396	0,000	Valid	Very High
10	0,603	0,396	0,017	Valid	High
11	0,349	0,396	0,203	Invalid	Low
12	0,986	0,396	0,000	Valid	Very High
13	0,703	0,396	0,003	Valid	High
14	0,986	0,396	0,000	Valid	Very High
15	0,851	0,396	0,000	Valid	Very High
16	0,986	0,396	0,000	Valid	Very High
17	0,986	0,396	0,000	Valid	Very High
18	0,986	0,396	0,000	Valid	Very High
19	0,258	0,396	0,352	Invalid	Low
20	0,986	0,396	0,000	Valid	Very High
21	0,267	0,396	0,336	Invalid	Low
22	0,986	0,396	0,000	Valid	Very High
23	0,737	0,396	0,002	Valid	High
24	0,114	0,396	0,687	Invalid	Very Low
25	0,986	0,396	0,000	Valid	Very High

Based on table 2 regarding the results of the test validity calculation using the IBM SPSS Statistics 25 application, it was found that there were 20 questions (80%) that could be said to be valid, namely numbers 1, 2, 3, 4, 5, 6, 7, 9, 10, 12, 13, 14, 15, 16, 17, 18, 20, 22, 23, 24, and 25. Then the remaining 5 questions (20%) were declared invalid, namely numbers 8, 11, 19, 21, and 24. This means that the majority of questions tested in the class have a valid test question category. Determination of the validity status of a test item is based on the calculated R

value being greater than the R Table or the significance value being less than 0.05 (Dahlan et al., 2024).

Reliability Test

Reliability testing is a tool to measure the consistency of an instrument used by researchers, ensuring that the instrument can be relied upon to measure the same questions repeatedly (Maulidiyah et al., 2020).

The reliability test in this study used IBM SPSS 25 using the Cronbach's alpha technique. The basis for the decisions made was partly as follows:

1. If the Cronbach's alpha value is >0.6, it is considered reliable.
2. If the Cronbach's alpha value is <0.6, it is considered less reliable.

Tabel 3. Cace Processing Summary

Reliability Statistics	
Cronbach's Alpha	N of Items
,966	25

Based on Table 3 above, the Cronbach's Alpha value was 0.966 with a total N of Items of 25, indicating the number of questions tested and answered by students. A Cronbach's Alpha value greater than 0.6 indicates the reliability of the test item.

Arikunto explained that the effectiveness of an instrument is determined by its validity and reliability. Instrument validity measures the accuracy of the intended measurement, while reliability measures the extent to which a measurement can be trusted (Mochammad Noor Akhmadi, 2021). An instrument is considered valid if it accurately describes the data from the variables without deviating from the actual situation. An instrument is considered reliable if it can reliably describe the data (Lestari et al., 2023).

Difficulty Level

According to Nana Sudjana, the assumption used to achieve good question quality, in addition to meeting validity and reliability, is a balance in the level of difficulty of the questions. This balance refers to the proportional distribution of easy, medium, and difficult questions (Okyranida et al., 2024).

In his book, Psychological Education, Witherington states that the level of difficulty of a learning outcome test item can be determined by the size of the number representing its difficulty. This number, which provides an indication of the item's difficulty level, is known as the difficulty index, which in the world of learning outcome evaluation is generally symbolized by the letter P, which stands for Proportion (Bramantha & Rahmania, 2022). A test should not be too easy, nor should it be too difficult. An item that is so easy that all students can

answer correctly is not a good item. Similarly, an item that is so difficult that students cannot answer it is also not good. Therefore, a good item is one that has a certain degree of difficulty (Virginia et al., 2021).

The difficulty level of a test item is a numerical representation of its difficulty. By comparing the total number of students who answered each item correctly with the number of students who answered each item correctly, the item's difficulty level can be determined. Questions are considered easier as the difficulty index approaches 1.00 (Mochammad Noor Akhmadi, 2021). To determine the difficulty level of a question, use the following formula:

$$\mathbf{P} = \frac{B}{JS}$$

Description:

P = difficulty index

B = number of students answering correctly

JS = total number of students

The interpretation of the level of difficulty of test items uses the criteria according to Witherington in Anas Sudijono as follows:

Table 4. *Item difficulty index/IKB*

INTERVAL	INTERPRETATION
0,00 – 0,30	Difficult
0,31 – 0,70	Medium
0,71 – 1,00	Easy

The level of difficulty of the questions obtained from the analysis of Arabic multiple choice questions in maharah qiro'ah using SPSS 25 is as follows:

Table 5. SPSS output results for Arabic language test item analysis

Question	Valid (N)	Missing	Mean
Question 1	15	0	0.6000
Question 2	15	0	0.6000
Question 3	15	0	0.6667
Question 4	15	0	0.6667
Question 5	15	0	0.6667
Question 6	15	0	0.6667
Question 7	15	0	0.7333
Question 8	15	0	0.7333
Question 9	15	0	0.6667
Question 10	15	0	0.7333
Question 11	15	0	0.8000

Question 12	15	0	0.6667
Question 13	15	0	0.6667
Question 14	15	0	0.6667
Question 15	15	0	0.7333
Question 16	15	0	0.6667
Question 17	15	0	0.6667
Question 18	15	0	0.6667
Question 19	15	0	0.8000
Question 20	15	0	0.6667
Question 21	15	0	0.7333
Question 22	15	0	0.6667
Question 23	15	0	0.6667
Question 24	15	0	0.8000
Question 25	15	0	0.6667

Tabel 6. Question item status

Question Status	Question Number	Total	Persentase
Easy	7, 8, 10, 11, 15, 19, 21, 24	8	32%
Medium	1, 2, 3, 4, 5, 6, 9, 12, 13, 14, 16, 17, 18, 20, 22, 23, 25	17	68%
Difficult	-	0	0%

From the table above, there are 8 questions on Arabic language subjects in maharah qiroah that have a difficulty level of between 0.71-1.00: the easy category if the percentage is 32%, meanwhile, 17 questions have a difficulty level between 0.31-0.70: the medium category or if the percentage is 32% and there are no questions that have a difficulty level between 0.00 - 0.30.

So it can be concluded that the items in maharah qiroah are seen from the level of difficulty, the medium category is more dominant, namely 68%, meaning the level of difficulty of the questions is not balanced. The results of this analysis illustrate that there are no questions in the difficult category. The comparison between easy-medium-difficult questions can be made 3-4-3, meaning 30% of the questions are in the easy category, 40% of the questions in the medium category and 30% of the questions in the difficult category. Another similar ratio is the 3-5-2 ratio, meaning 30% of the questions are easy, 50% are medium, and 20% are difficult (Khofifah et al., 2020).

A good question is neither too easy nor too difficult. Students will not improve their problem-solving skills if the question is too easy. Conversely, if the question is too difficult, students will lose hope and motivation to try again because it seems impossible to achieve (Ropii & Fahrurrozi, 2017).

Differential Power

Discriminatory power is the ability of an item to differentiate between high-ability students and low-ability students. Discriminatory power is calculated by subtracting the proportion of participants in the upper group who answered

correctly from the proportion of participants in the lower group who answered correctly (Purniasari et al., 2021). Hendriana and Soemarmo also stated that a test item is said to have discriminatory power if it can differentiate the quality of answers between students who understand and those who do not (Okyranida et al., 2024). The higher the discriminatory power, the better the item is at differentiating between high-ability students and low-ability students.

In this study, the researchers first calculated the total scores of students, sorted them from highest to lowest, determined the upper and lower groups, and calculated the average scores of the upper and lower groups. The results are interpreted using Arikunto's four discriminatory power criteria:

Table 7. *Distinguishing Power Category*

DP	Category
0,00 - 0,19	Poor
0,20 - 0,39	Fair
0,40 - 0,69	Good
0,70 - 1,00	Very Good

Effectiveness of Deception

Distractor effectiveness analysis, or response pattern analysis, is conducted by counting the number of test participants who choose each answer alternative for each test item. A distractor is considered good if it is chosen by 5% of test participants, and is further grouped into two criteria: effectively functioning distractors and ineffective distractors (Halik et al., 2019). To interpret the distractor effectiveness for each test item, the following criteria are used, adapted from the Likert scale (Purniasari et al., 2021):

The effectiveness of the distractor is said to be very good if all four distractors function.

The effectiveness of the distractor is said to be good if three distractors function.

The effectiveness of the distractor is said to be quite good if two distractors function.

The effectiveness of the distractor is said to be poor if one distractor functions.

The effectiveness of the distractor is said to be poor if all distractors do not function.

Calculation results for the effectiveness of distractors paraphrased in percentage form for each question item against the answer choices as follows:

Table 9. The distractor works effectively and the distractor does not work effectively.

NO	DISTRACTOR EFFECTIVENESS				DESCRIPTION
	A	B	C	D	
1	0%	0%	100%	0%	All distractors are poor
2	0%	100%	0%	0%	All distractors are poor
3	0%	100%	0%	0%	All distractors are poor
4	13%	0%	67%	20%	Distractor b is poor
5	13%	0%	20%	67%	Distractor b is poor
6	7%	13%	67%	13%	All distractors are good
7	0%	0%	0%	100%	All distractors are poor
8	0%	0%	100%	0%	All distractors are poor
9	67%	0%	13%	20%	Distractor b is poor
10	80%	13%	7%	0%	Distractor d is poor
11	0%	93%	7%	0%	Distractors a and d are poor Good
12	0%	0%	0%	100%	All distractors are poor
13	33%	60%	0%	7%	Distractor c is not good
14	0%	100%	0%	0%	All distractors are poor
15	0%	0%	100%	0%	All distractors are poor
16	0%	100%	0%	0%	All distractors are poor
17	73%	7%	20%	0%	Distractor d is poor
18	87%	13%	0%	0%	Distractors c and d are not good
19	100%	0%	0%	0%	All distractors are poor
20	0%	100%	0%	0%	All distractors are poor
21	7%	0%	13%	80%	Distractor b is poor
22	13%	0%	73%	13%	Distractor b is poor
23	7%	0%	0%	93%	Distractors b and c are not good Good
24	7%	93%	0%	0%	Distractors c and d are not good
25	67%	20%	0%	13%	Distractor b is poor

The table above shows that the distracting power of the PAS items is 24 out of 25, which are classified as poor distractors due to the multiple-choice answer percentage being less than 5%. Meanwhile, only one question is categorized as good because all multiple-choice answers have a percentage above 5%. Therefore, the majority of questions have distracting power that requires better replacement questions (Wirandani et al., 2019)

CONCLUSION

The findings of this study demonstrate that most test items administered to class X IPA MA Al-Rosyid Bojonegoro fall into the valid category, as indicated by r-count values exceeding r-table and significance levels below 0.05. The reliability analysis further shows a strong Cronbach's Alpha value of 0.966, confirming that the 25-item test is highly consistent. However, the distribution of difficulty levels is unbalanced, with 68% of items classified as medium and the absence of items in the difficult category. Ideal proportions such as 3-4-3 or 3-5-2 could offer a more optimal distribution across easy, medium, and difficult items. Additionally, the discrimination index reveals that 72% of items fall into the poor category, and distractor effectiveness is weak, with 24 items showing distractors below the 5% threshold. These findings highlight the need for more rigorous test construction to improve the accuracy and fairness of student assessment.

Despite providing a comprehensive analysis of test validity, reliability, item difficulty, discrimination power, and distractor effectiveness, this study is limited by its focus on a single school and a relatively small sample size. Its scholarly contribution lies in offering a detailed empirical evaluation of teacher-made tests, which can serve as a practical reference for improving item development in educational settings. Future research should involve larger and more diverse samples, comparative multi-school analyses, and the integration of digital test development tools to enhance generalizability and produce more robust guidelines for constructing high-quality assessment instruments.

REFERENCES

Abu Hasan, Agus R., & Maripaz C. Abas, N. D. K. (2025). Improving Brand Awareness of Educational Institutions Through Educational Personnel Recruitment Management in Madrasah. At-Tarbiyat, 8.

Aini, T. N., & Wahid, A. H. (2024). Psychological Strategies for Building Quality Human Resources in Madrasah. Proceeding of International Conference on Education, Society and Humanity, 2(1), 154–160.

Alam, A., & Mohanty, A. (2023). Cultural Beliefs and Equity in Educational Institutions: Exploring the Social and Philosophical Notions of Ability Groupings in Teaching and Learning of Mathematics. International Journal of Adolescence and Youth, 28(1), 2270662. https://doi.org/10.1080/02673843.2023.2270662

Barokah, M. (2025). Management of Learning Outcomes Through SIJAGU PAI: Design and Implementation of a Digital Reporting System for Islamic Religious Education. Journal of Educational Management Research, 4(2), 845–860.

Dewi, A. T. A., Najiburohman, & Hefniy. (2025). Virtual School Tours: Boosting Community Interest and Attracting Prospective Students. Evaluasi: Jurnal Manajemen Pendidikan Islam, 9(2), 340–353.

Fajariyah, H. (2025). Self Directed Learning: Meningkatkan Kepercayaan Diri dalam Berbicara Bahasa Arab. Pendas: Jurnal Ilmiah Pendidikan Dasar, 10(1), 339–353.

Hadi, N., & Masuwd, M. A. (2025). Classical Cooperative Learning Model for Reading Classic Literature: Enhancing Student Independence Through Self-Regulation. Izdihar: Journal of Arabic Language Teaching, Linguistics, and Literature, 8(1). https://doi.org/10.22219/jiz.v8i1.36829

Indayati, T. (2025). Pengembangan Kemampuan Lisan Berbahasa Arab: Integrasi Operant Conditioning dalam Lingkungan Pembelajaran Bahasa yang Holistik. Pendas: Jurnal Ilmiah Pendidikan Dasar, 10(1), 221–234.

Jamil, T. I., & Sanusi, S. F. (2024). Enhancing Student Learning Outcomes in PAI Subjects: The Impact of PowerPoint Learning Media Application. Educazione: Journal of Education and Learning, 1(2), 66–77. https://doi.org/10.61987/educazione.v1i2.502

Jannah, F., & Rizquha, A. (2025). Deconstructing Dogmatic Narratives: An Effort to Recontextualize Islamic Education Material for the Critical Generation. Jurnal Islam Nusantara, 9(1), 43–56.

Khoiroh, U. (2025). Emotional Management in Local Wisdom: Strategies for Enhancing Teachers' Work Resilience in Pesantren-Based Madrasah. Journal of Educational Management Research, 4(5), 2296–2309.

Kholifatunnisak, K. (2024). Manajemen Kurikulum PAI dalam Meningkatkan Karakter Religius Peserta Didik di MTs Azzainiyah 1 Randumerak. Jurnal Educatio FKIP UNMA, 10(3). https://doi.org/10.31949/educatio.v10i3.9122

Kusnanto, N., Sukristyanto, A., & Rochim, A. I. (2023). Relevance of National Education Policies as an Effort to Improve the Quality of Madrasah Tsanawiyah Education Services. The Spirit of Society Journal: International Journal of Society Development and Engagement, 6(2), 136–151. https://doi.org/10.29138/scj.v6i2.2210

Ma'isyah, M., Rizal, M. S., Iqna'a, F. J., & Setiawan, B. A. (2024). Dynamics of Islamic Boarding Schools in Facing Globalization: Integration Between Tradition and Modernity. Proceeding of International Conference on Education, Society and Humanity, 2(2), 71–80.

Makiyah, N. (2024). Enhancing Educational Excellence: Elevating Learning Quality Through Podcast-Based Arts Performances in Pesantren. Journal of Islamic Education Research, 5(1), 1–12. https://doi.org/10.35719/jier.v5i1.371

Manshur, U., Rozi, F., Saleha, L., & Sholihah, C. (2024). Eco-Friendly Media: Assistance in Developing Educational Props From Waste Materials in Probolinggo City. GUYUB: Journal of Community Engagement, 5(1), 249–271. https://doi.org/10.33650/guyub.v5i1.8317

Muharromah, L. (2025). Digital Ethics in the Perspective of Islamic Education: Cultivating Religious Awareness in Cyberspace. Journal of Educational Management Research, 4(3), 1280–1293.

Mukarromah, A., Manshur, U., & Syafaat, I. N. (2024). Between Tradition and Modernity: The Relevance of Boarding School Values in Forming Students' Obedient Behavior. Proceeding of International Conference on Education, Society and Humanity, 2(2), 767–774.

Nuriyah, K., Putri, D. M. S., & Anisa, Z. (2024). Penerapan Prinsip Psikologi Positif dalam Kebijakan Manajemen Sumber Daya Manusia untuk Meningkatkan Kinerja dan Kepuasan Kerja. Jurnal Educatio FKIP UNMA, 10(2), 687–694.

Yakin, A. (2025). Transforming Organizational Culture in Islamic Educational Institutions: Cultivating a Quality-Oriented Learning Environment for Academic Excellence. Journal of Educational Management Research, 4(4), 1711–1731.

Zakiyah, L. (2025). Mitigating Adolescent Mythomania Through Kindness-Based Pedagogy: Lesson Learned From Rural Schools. Cendekia: Jurnal Kependidikan dan Kemasyarakatan, 23(2).

Abu Hasan, Agus R., & Abas, M. C. M. N. D. K. (2025). Improving Brand Awareness of Educational Institutions Through Educational Personnel Recruitment Management in Madrasah. At-Tarbiyat, 8.

Alam, A., & Mohanty, A. (2023). Cultural Beliefs and Equity in Educational Institutions: Exploring the Social and Philosophical Notions of Ability Groupings in Teaching and Learning of Mathematics. International Journal of Adolescence and Youth, 28(1), 2270662. https://doi.org/10.1080/02673843.2023.2270662

Kusnanto, N., Sukristyanto, A., & Rochim, A. I. (2023). Relevance of National Education Policies as an Effort to Improve the Quality of Madrasah Tsanawiyah Education Services. The Spirit of Society Journal: International Journal of Society Development and Engagement, 6(2), 136–151. https://doi.org/10.29138/scj.v6i2.2210

Kwek, D., Ho, J., & Wong, H. M. (2023). Singapore's Educational Reforms Toward Holistic Outcomes: (Un)intended Consequences of Policy Layering. Center for Universal Education at The Brookings Institution (Case Study).

Safarova, R. (2023). Computer-Didactic Support for the Training of Social Sphere Specialists at the University Based on a Cultural Approach. Problems in the Textile and Light Industry in the Context of Integration of Science and Industry and Ways to Solve Them (PTLICISIWS-2022), 2789(1), 050009. https://doi.org/10.1063/5.0149686.