AL-ADABIYAH: Jurnal Pendidikan Agama Islam
http://al-adabiyah.uinkhas.ac.id/

Evaluating the Quality of Islamic Religious Education Examination Items through Classical Test Theory Analysis

Rohanis Aulyah,[1*] Analis Realita Shohih,[2] Radha Cantika Fitra,[3] Irfan,[4] Irfandi,[5] Aditya Nugroho,[6] Ubaydillah[7]
[1-6] Institut Agama Islam Negeri Kendari, Indonesia
[*]rohanisaulyah@gmail.com

Abstract

The quality of examination items plays a crucial role in ensuring the accuracy and fairness of educational assessment. This study aimed to evaluate the quality of Islamic Religious Education (PAI) Final Semester Examination items for Grade X students at SMK Negeri 3 Kolaka Timur using the Classical Test Theory (CTT) framework. Specifically, the study examined item validity, reliability, difficulty level, and discrimination power to determine the effectiveness of the examination instrument in measuring student achievement. A descriptive quantitative research design was employed using data derived from 23 students' answer sheets, examination blueprints, answer keys, and supporting interview data from the subject teacher. Data were analyzed through Pearson Product-Moment correlation for item validity, Cronbach's Alpha for reliability, difficulty index analysis, and discrimination index analysis. The findings revealed that 14 out of 20 items (70%) met the validity criteria, while 6 items (30%) were classified as invalid. The instrument demonstrated high reliability with a Cronbach's Alpha coefficient of 0.828, indicating strong internal consistency. However, the difficulty level analysis showed an imbalanced distribution of items, with 65% categorized as easy and 35% as moderate, while no difficult items were identified. Furthermore, discrimination analysis indicated that 75% of the items exhibited poor discrimination power and only 25% achieved a fair category. These results suggest that although the instrument was reliable, it was less effective in differentiating students according to their achievement levels. The novelty of this study lies in its comprehensive application of Classical Test Theory to evaluate Islamic Religious Education examination items within a vocational secondary school context, an area that remains underexplored in assessment research. This study contributes to the literature on educational measurement by demonstrating that high reliability alone does not guarantee overall test quality and highlighting the importance of integrating validity, difficulty level, and discrimination analyses in the development of evidence-based assessment practices.

Keywords: Classical Test Theory, Educational Assessment, Islamic Religious Education, Item Analysis, Reliability, Validity

Introduction

Assessment constitutes an integral component of the educational process, serving not only to measure students' competency achievement but also to evaluate the effectiveness of instructional practices and curriculum implementation (B., 2017). In contemporary educational systems, assessment functions as a mechanism for gathering evidence about student learning and informing educational decision-making at both classroom and institutional levels. Effective assessment provides meaningful feedback for students, teachers, and policymakers regarding the extent to which learning objectives have been achieved (Black & Wiliam, 2018). Consequently, the quality of assessment instruments has become a critical concern in educational measurement because inaccurate instruments may lead to misleading interpretations of student achievement and educational quality.

Among the various forms of assessment employed in schools, the Final Semester Examination (*Ujian Akhir Semester/UAS*) plays a particularly important role in determining students' mastery of learning materials accumulated throughout an academic semester (Wahyuni & Dinisa Putri, 2020). The results of semester examinations are frequently used as indicators of learning success, promotion decisions, and curriculum effectiveness. Therefore, examination items must be carefully designed to ensure that they accurately measure the intended competencies. Poorly constructed test items may fail to capture students' actual abilities and potentially undermine the validity of educational decisions based on examination results.

The quality of a test instrument fundamentally determines the accuracy and credibility of assessment outcomes. According to educational measurement theory, a high-quality test should satisfy several psychometric requirements, including validity, reliability, appropriate difficulty level, and adequate discrimination power (Yulianti et al., 2018). Validity refers to the extent to which a test measures the construct it intends to measure, while reliability indicates the consistency of measurement results across administrations (DeVellis & Thorpe, 2021). Difficulty level reflects the proportion of students who correctly answer an item, whereas discrimination power indicates the ability of an item to distinguish between high-performing and low-performing students (Haladyna & Rodriguez, 2013). These indicators collectively provide comprehensive information regarding the effectiveness of assessment instruments in measuring student achievement.

One of the most widely used frameworks for evaluating test quality is Classical Test Theory (CTT). CTT has been extensively applied in educational assessment because of its practical applicability and relatively simple analytical procedures (Crocker & Algina, 2008). Through item analysis, educators can identify problematic questions, examine item performance, and revise test instruments based on empirical evidence. Studies have consistently demonstrated that systematic item analysis contributes significantly to improving assessment quality and enhancing the accuracy of educational measurement (Tosuncuoglu, 2018). Furthermore, item analysis supports the

development of item banks and promotes evidence-based assessment practices within educational institutions.

Previous studies have investigated the quality of educational assessment instruments across various subject areas and educational levels. Research conducted by Putri et al. (2023) found that biology examination items were predominantly categorized as easy and exhibited limited discriminatory power. Similarly, Prabowo (2024) reported that several Islamic Religious Education (PAI) assessment items failed to meet validity standards, indicating weaknesses in test construction procedures. Other studies have also highlighted recurring issues related to item validity, distractor effectiveness, and reliability in school-based assessments (Azizah & Supahar, 2024; Saputri et al., 2023). Despite these findings, many schools continue to implement examinations without conducting systematic psychometric evaluations of their assessment instruments.

Although studies on item analysis have been widely conducted, several gaps remain in the literature. First, empirical studies examining the psychometric quality of Islamic Religious Education examination items at the vocational high school (SMK) level remain limited compared to studies in science and mathematics education. Second, existing studies often focus only on selected psychometric indicators, such as validity or reliability, without providing a comprehensive analysis encompassing validity, reliability, difficulty level, and discrimination power simultaneously. Third, evidence regarding the quality of PAI assessment instruments in vocational education contexts, particularly in eastern regions of Indonesia, is still scarce. These limitations indicate the need for further research that provides a comprehensive evaluation of PAI examination instruments using established measurement frameworks.

The importance of such research is particularly evident in Islamic Religious Education, which aims not only to develop students' cognitive understanding of Islamic teachings but also to foster affective and behavioral competencies. Consequently, assessment instruments used in PAI must be capable of accurately measuring learning outcomes and distinguishing varying levels of student mastery. Inadequate assessment quality may lead to inaccurate conclusions regarding students' religious knowledge and competencies, ultimately affecting instructional improvement efforts.

Based on these considerations, this study seeks to evaluate the quality of Final Semester Examination items in Islamic Religious Education for Grade X students at SMK Negeri 3 Kolaka Timur through the lens of Classical Test Theory. Specifically, the study analyzes item validity, reliability, difficulty level, and discrimination power to determine the overall quality of the examination instrument. The findings are expected to contribute to the improvement of assessment practices in Islamic Religious Education, provide empirical evidence for the development of higher-quality test instruments, and strengthen data-driven educational evaluation in vocational secondary education.

Method

This study employed a descriptive quantitative research design to evaluate the psychometric quality of Islamic Religious Education (PAI) Final Semester Examination (UAS) items using the Classical Test Theory (CTT) framework. A descriptive quantitative approach was selected because it enables researchers to systematically describe, quantify, and interpret the quality characteristics of test items based on empirical numerical evidence derived from student responses (Jayusman & Shavab, 2020). In educational measurement research, descriptive quantitative methods are widely used to examine item performance indicators, including validity, reliability, difficulty level, and discrimination power, which collectively provide comprehensive information regarding the effectiveness of assessment instruments (Crocker & Algina, 2008).

The study was conducted at SMK Negeri 3 Kolaka Timur, Southeast Sulawesi, Indonesia. The research focused on the Final Semester Examination administered in the Islamic Religious Education subject for Grade X students during the 2025/2026 academic year. The primary data consisted of answer sheets from 23 students who completed the examination. Although the number of participants was relatively limited, the entire population of available Grade X student responses was included in the analysis. In Classical Test Theory, item analysis can still be conducted with limited samples for preliminary evaluation purposes, particularly when the objective is to assess the quality of classroom-based examinations and generate practical recommendations for test improvement (Haladyna & Rodriguez, 2013). Therefore, the present study is positioned as a school-level evaluation intended to provide empirical feedback for improving assessment practices rather than generalizing findings to broader populations.

Data were collected through documentation and semi-structured interview techniques. Documentation served as the primary source of quantitative data and included students' answer sheets, examination blueprints, test specifications, answer keys, and scoring guidelines. These documents provided the basis for calculating psychometric indices and evaluating item quality. In addition, semi-structured interviews were conducted with the Islamic Religious Education teacher responsible for constructing the examination, Mrs. Sri Puji Lestari, S.Pd.I. The interviews aimed to obtain contextual information regarding item development procedures, blueprint preparation, validation practices, alignment between learning objectives and test items, and challenges encountered during examination construction. The integration of documentary evidence and interview data allowed for a more comprehensive interpretation of the quantitative findings and strengthened the credibility of the study.

The study adopted Classical Test Theory (CTT) as the analytical framework. CTT remains one of the most frequently applied approaches in educational assessment because of its practicality, straightforward computational procedures, and effectiveness in evaluating classroom-based

examinations (Crocker & Algina, 2008; DeVellis & Thorpe, 2021). Within this framework, observed scores are assumed to consist of true scores and measurement error, making item-level analysis essential for identifying weaknesses in test instruments. Previous studies have demonstrated that CTT-based item analysis provides valuable information for improving educational assessments, developing item banks, and enhancing the accuracy of measurement outcomes (Tosuncuoglu, 2018).

The quantitative data were analyzed using Microsoft Excel and SPSS statistical software. Student responses were coded dichotomously, assigning a score of 1 for correct answers and 0 for incorrect answers. Subsequently, four psychometric indicators were examined, namely item validity, reliability, difficulty level, and discrimination power. These indicators were selected because they represent the fundamental criteria for evaluating educational assessment quality and are commonly employed in classroom-based measurement studies (Hermanto et al., 2021).

Item validity was examined using the Pearson Product-Moment correlation coefficient, which measures the relationship between individual item scores and total test scores (Arikunto, 2013). The validity coefficient was calculated using the following formula:

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{(N\sum X^2 - (\sum X)^2)(N\sum Y^2 - (\sum Y)^2)}}$$

Items were classified as valid when the obtained correlation coefficient (r_calculated) exceeded the critical value of r_table at a significance level of 0.05. With a sample size of 23 students, the critical value of r_table was 0.413 (Saputri et al., 2023). Valid items indicate that the question measures the same construct represented by the overall test, whereas invalid items suggest inconsistencies between item content and the intended learning objectives.

Instrument reliability was assessed using Cronbach's Alpha coefficient, which measures the internal consistency of test items and indicates the extent to which items function cohesively as a measurement instrument (Naga, 2004). Cronbach's Alpha was calculated using the following formula:

$$r_{ii} = \left[\frac{k}{k-1}\right]\left(1 - \frac{\sum S_i^2}{S_t^2}\right)$$

where k represents the number of items, ΣSi^2 represents the sum of item variances, and St^2 denotes total score variance. Reliability interpretation followed the criteria proposed by Yusup (2017): very high (0.80–1.00), high (0.60–0.79), moderate (0.40–0.59), low (0.20–0.39), and very low (0.00–0.19). Reliability analysis is important because a test cannot produce meaningful assessment results if its scores are unstable or inconsistent across measurements.

The difficulty level of each item was determined using the difficulty index (P), calculated as:

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

where B represents the number of students answering the item correctly and JS represents the total number of students taking the test. The interpretation criteria followed Lestari et al. (2023): difficult (0.00–0.30), moderate (0.31–0.70), and easy (0.71–1.00). Difficulty analysis provides information regarding the extent to which an item challenges students and helps determine whether the examination contains a balanced distribution of easy, moderate, and difficult questions.

Discrimination power was analyzed to determine the ability of each item to distinguish between high-performing and low-performing students. Following Classical Test Theory procedures, students were ranked according to their total test scores and divided into upper and lower groups based on achievement levels. The discrimination index (D) was calculated using the following formula:

$$\mathbf{D} = \frac{\mathbf{JA}}{\mathbf{BA}} - \frac{\mathbf{JB}}{\mathbf{BB}}$$

where BA denotes the number of correct responses in the upper group, BB denotes the number of correct responses in the lower group, JA denotes the number of students in the upper group, and JB denotes the number of students in the lower group (Solichin, 2017). The interpretation criteria were as follows: D < 0.20 (poor), 0.20–0.39 (fair), 0.40–0.69 (good), and 0.70–1.00 (very good). Items with negative discrimination indices were considered defective and recommended for revision or elimination. Discrimination analysis is particularly important because it indicates the effectiveness of an item in differentiating varying levels of student competence (Haladyna & Rodriguez, 2013).

To enhance the trustworthiness of the findings, data verification procedures were conducted through cross-checking answer sheets, answer keys, and scoring records before statistical analysis. Interview findings were also used to support the interpretation of quantitative results, particularly when explaining the causes of invalid items or weak discrimination indices. Through these procedures, the study sought to provide a comprehensive and evidence-based evaluation of the quality of Islamic Religious Education examination items at the vocational secondary school level.

Results and Discussion

Results

Item Validity Analysis

The validity of the 20 Islamic Religious Education (PAI) Final Semester Examination items was examined using the Pearson Product-Moment correlation coefficient. The analysis was conducted by correlating individual item scores with the total test score obtained by each student. An item was considered valid when its calculated correlation coefficient (*r*_{calculated}) exceeded the critical value of *r*_{table} (0.413) at the 5% significance level. The results revealed that 14 out of 20 items (70%) met the validity criterion, whereas 6 items (30%) failed to reach the required threshold. These findings indicate that most examination items were able to measure the intended construct consistently with the overall test, although several items demonstrated insufficient correlations with students' total achievement scores.

Table 1. Item Validity Test Results

Item No.	r_Calculated	r_Table (N=23, α=5%)	Status
1	0.492	0.413	Valid
2	0.346	0.413	Invalid
3	0.488	0.413	Valid
4	0.302	0.413	Invalid
5	0.469	0.413	Valid
6	0.596	0.413	Valid
7	0.453	0.413	Valid
8	0.777	0.413	Valid
9	0.760	0.413	Valid
10	0.453	0.413	Valid
11	0.433	0.413	Valid
12	0.542	0.413	Valid
13	0.263	0.413	Invalid
14	0.325	0.413	Invalid
15	0.550	0.413	Valid
16	0.731	0.413	Valid
17	0.200	0.413	Invalid
18	0.550	0.413	Valid
19	0.456	0.413	Valid
20	0.604	0.413	Valid

Source: Primary Data Processed (2026)

Further examination of the validity coefficients shows substantial variation among items. The highest validity coefficient was observed in Item 8 (r = 0.777), followed by Item 9 (r = 0.760) and Item 16 (r = 0.731), indicating strong relationships between these items and overall student performance. Conversely, Items 17 (r = 0.200), 13 (r = 0.263), and 4 (r = 0.302) demonstrated weak correlations with total test scores, suggesting that these items may not adequately represent the competencies measured by the examination. Overall, the findings indicate that while the majority of items demonstrated acceptable

validity, several items require revision before being reused in future assessments.

Instrument Reliability Analysis

The internal consistency of the examination instrument was evaluated using Cronbach's Alpha coefficient. The analysis produced a reliability coefficient of α = 0.828, indicating a high level of reliability according to the classification proposed by Yusup (2017). This result suggests that the set of examination items functioned cohesively in measuring students' achievement in Islamic Religious Education.

A reliability coefficient exceeding 0.80 indicates that the instrument produced relatively stable and consistent scores across the examined items. In practical terms, students who performed well on certain items tended to perform consistently well on other items measuring similar competencies. The high reliability coefficient further suggests that measurement error within the instrument was relatively low. Therefore, the examination can be considered sufficiently reliable for evaluating student learning outcomes at the classroom level.

Nevertheless, reliability should not be interpreted as the sole indicator of test quality. Although the instrument demonstrated strong internal consistency, item-level analyses revealed that several questions did not satisfy validity requirements and many items exhibited weak discrimination power. These findings indicate that a reliable instrument may still contain items requiring improvement to enhance overall assessment quality.

Difficulty Level Analysis

The difficulty level analysis was conducted to determine the proportion of students who answered each item correctly. Based on the calculated difficulty indices, the majority of items were categorized as easy. Of the 20 examination items analyzed, 13 items (65%) fell into the easy category, while 7 items (35%) were classified as moderate. No items were categorized as difficult.

Table 2. Distribution of Difficulty Levels

Category	P Range	Number of Items	Percentage
Easy	0.71–1.00	13	65%
Moderate	0.31–0.70	7	35%
Difficult	0.00–0.30	0	0%
Total		**20**	**100%**

Source: Primary Data Processed (2026)

The distribution presented in Table 2 demonstrates a considerable concentration of items within the easy category. More than half of the examination questions could be answered correctly by most students, indicating that the test may not have provided sufficient challenge for higher-achieving learners. At the same time, the absence of difficult items suggests a limited ability of the examination to assess higher-order cognitive competencies or advanced mastery of learning content.

The dominance of easy items also affects the overall measurement quality of the test. When a large proportion of students answer the same items correctly, the examination becomes less sensitive to differences in student achievement levels. Consequently, although the test may effectively measure basic understanding, its capacity to distinguish varying levels of mastery may be reduced.

Discrimination Power Analysis

Discrimination power analysis was performed to evaluate each item's ability to differentiate between students with high and low overall achievement. The results revealed that most items possessed weak discriminatory capacity. Out of the 20 items analyzed, 15 items (75%) were categorized as poor, while only 5 items (25%) achieved the fair category. No items reached the good or very good classification.

Table 3. Discrimination Power Analysis Results

Item No.	Discrimination Index (D)	Category
1	0.09	Poor
2	0.09	Poor
3	0.18	Poor
4	0.27	Fair
5	0.09	Poor
6	0.14	Poor
7	0.32	Fair
8	0.18	Poor
9	0.23	Fair
10	0.05	Poor
11	0.14	Poor
12	0.18	Poor
13	0.14	Poor
14	0.18	Poor
15	0.32	Fair
16	0.27	Fair
17	0.09	Poor
18	0.14	Poor
19	0.14	Poor
20	0.18	Poor

Source: Primary Data Processed (2026)

The discrimination indices ranged from 0.05 to 0.32, indicating generally weak performance across the examination items. Item 10 exhibited the lowest discrimination index (D = 0.05), suggesting that it provided minimal differentiation between high- and low-performing students. In contrast, Items 7 and 15 obtained the highest discrimination indices (D = 0.32), although they remained within the fair category and did not reach the threshold for good discrimination.

Overall, the results indicate that the examination items were relatively ineffective in distinguishing students with different levels of academic achievement. The predominance of poor discrimination indices suggests that

many items may have been too easy, insufficiently challenging, or inadequately constructed. Consequently, the examination's capacity to accurately classify students according to their actual competence levels was limited. These findings highlight the need for systematic item revision and more rigorous test development procedures in future administrations.

Discussion

The findings of this study indicate that the Islamic Religious Education (PAI) Final Semester Examination instrument demonstrated a high level of reliability (α = 0.828), suggesting that the test items functioned consistently in measuring student achievement. According to Classical Test Theory (CTT), reliability reflects the degree to which an instrument produces stable and consistent measurements across items and respondents (Crocker & Algina, 2008). A reliability coefficient exceeding 0.80 is generally considered indicative of strong internal consistency and acceptable measurement precision (DeVellis & Thorpe, 2021). However, reliability alone does not guarantee the overall quality of a test instrument. Educational measurement scholars emphasize that a reliable instrument may still fail to accurately assess the intended construct if weaknesses exist in item validity and discrimination power (Haladyna & Rodriguez, 2013). This phenomenon appears evident in the present study, where the instrument demonstrated high reliability despite several items exhibiting inadequate validity and poor discriminatory capacity. Consequently, the findings reinforce the notion that reliability should be interpreted alongside other psychometric indicators rather than as a standalone measure of assessment quality.

The validity analysis revealed that several examination items failed to satisfy the established validity criterion, indicating that they were not sufficiently aligned with the overall construct measured by the test. Specifically, Items 2, 4, 13, 14, and 17 produced correlation coefficients below the critical threshold, suggesting that these items contributed minimally to the measurement of student competence. From a psychometric perspective, low item validity often emerges when questions do not adequately represent learning objectives, contain ambiguous wording, or include distractors that fail to function as intended (Azizah & Supahar, 2024). Interview data further indicated that the examination items were developed primarily based on teaching experience and practical classroom considerations rather than through systematic validation procedures. This finding is consistent with previous studies reporting that many classroom-based assessments are developed without formal expert review, empirical pilot testing, or psychometric evaluation (Brookhart, 2018). As a result, certain items may measure peripheral knowledge rather than the competencies specified in the assessment blueprint. The present findings support recommendations from assessment literature emphasizing that content validity should be established through rigorous alignment between learning outcomes, instructional objectives, and assessment tasks before examinations are administered (Nitko & Brookhart, 2014).

Another important finding concerns the distribution of item difficulty levels. The analysis demonstrated that 65% of the items were categorized as easy, while no items were classified as difficult. This distribution suggests that the examination predominantly assessed lower-level cognitive achievement and may not have adequately challenged students with higher levels of competence. According to educational assessment principles, an effective examination should contain a balanced range of item difficulties to capture variations in student ability and provide meaningful information across the achievement spectrum (Haladyna & Rodriguez, 2013). Excessive concentrations of easy items can reduce score variability because most students answer correctly, thereby limiting the instrument's capacity to differentiate levels of mastery. Similar findings have been reported by Putri et al. (2023), who found that Biology examination items were dominated by easy questions, resulting in limited measurement sensitivity. In the context of Islamic Religious Education, the predominance of easy items may indicate an overemphasis on factual recall and memorization rather than higher-order thinking skills such as analysis, evaluation, and application of religious concepts. This observation is particularly important because contemporary educational frameworks increasingly emphasize the development of critical thinking, problem-solving, and reflective reasoning competencies (Anderson & Krathwohl, 2001).

The most critical issue identified in this study was the generally weak discrimination power of the examination items. Approximately 75% of the items were classified as poor discriminators, and none achieved the good or very good categories. Within Classical Test Theory, discrimination power is considered one of the most important indicators of item quality because it reflects an item's ability to distinguish between students with different levels of achievement (Ebel & Frisbie, 1991). Items with low discrimination indices contribute little to meaningful measurement because both high-performing and low-performing students tend to respond similarly. The findings of this study suggest that many items failed to differentiate students according to their actual competence levels, thereby reducing the effectiveness of the examination as an evaluative instrument. Similar results were reported by Imam (2023), who found that numerous Islamic Religious Education assessment items were rejected because of poor discrimination performance. Educational measurement literature suggests that low discrimination frequently occurs when items are excessively easy, excessively difficult, ambiguously worded, or supported by ineffective distractors (Rodriguez, 2016). Given that the majority of items in the present study were categorized as easy, the weak discrimination indices observed are likely a direct consequence of the limited variability in student responses.

The relationship between difficulty level and discrimination power warrants particular attention. Several studies have demonstrated that items with moderate difficulty levels generally exhibit stronger discrimination because they allow greater variation in student responses (Tarrant et al., 2009). Conversely, very easy items tend to generate uniformly correct

responses, thereby limiting their ability to distinguish between students with different levels of achievement. This pattern appears consistent with the results of the present study, where the predominance of easy items coincided with generally poor discrimination indices. Consequently, improving discrimination power may require not only revising poorly performing items but also creating a more balanced distribution of item difficulties. Teachers should be encouraged to construct assessment items representing a range of cognitive complexity levels, including analytical and evaluative tasks that require students to demonstrate deeper conceptual understanding.

The findings also highlight the importance of distractor quality in multiple-choice examinations. Although distractor effectiveness was not directly analyzed in this study, interview findings and psychometric results suggest that ineffective distractors may have contributed to the weak performance of several items. Research has consistently shown that distractors play a critical role in determining item validity and discrimination power because they influence students' response patterns and reduce the likelihood of random guessing (Rodriguez, 2005). Anshari et al. (2024) similarly emphasized that distractor analysis should be integrated into routine item evaluation procedures to improve examination quality. Future assessment development efforts should therefore include systematic reviews of distractor performance alongside traditional item analysis indicators.

From a practical perspective, the findings reveal that assessment development practices at SMK Negeri 3 Kolaka Timur have not yet fully incorporated evidence-based measurement principles. Although teachers routinely prepare examinations, item analysis results indicate that psychometric evaluation is not systematically conducted before or after test administration. This situation is not unique to the present context and has been reported in numerous educational settings where assessment practices remain largely experience-based rather than data-driven (Brookhart, 2018). The absence of systematic item evaluation may lead to the repeated use of low-quality questions, reducing the accuracy and fairness of educational assessment. Therefore, schools should establish regular assessment review procedures, including item analysis, expert validation, and item bank development, as part of institutional quality assurance initiatives.

The present study contributes to the growing body of literature on educational assessment within Islamic Religious Education by providing empirical evidence regarding the psychometric characteristics of examination items at the vocational secondary school level. While previous studies have examined item quality in various subject areas, research focusing specifically on PAI assessments in vocational education remains relatively limited. The findings demonstrate that high reliability does not necessarily imply overall test quality and underscore the need for comprehensive psychometric evaluation encompassing validity, reliability, difficulty level, and discrimination power simultaneously. These insights may assist educators, curriculum developers, and school administrators in designing more effective assessment systems that accurately measure student learning outcomes.

Despite its contributions, several limitations should be acknowledged. The study involved a relatively small sample of 23 students drawn from a single vocational school, limiting the generalizability of the findings. Furthermore, the analysis was confined to Classical Test Theory indicators and did not examine distractor effectiveness or employ more advanced approaches such as Item Response Theory (IRT). Future studies should involve larger and more diverse samples, incorporate distractor analysis, and compare CTT and IRT approaches to obtain a more comprehensive understanding of assessment quality in Islamic Religious Education.

Conclusion

This study evaluated the quality of Islamic Religious Education (PAI) Final Semester Examination items for Grade X students at SMK Negeri 3 Kolaka Timur using the Classical Test Theory (CTT) framework. The findings revealed that although the examination instrument demonstrated high reliability ($\alpha = 0.828$), significant weaknesses remained in several psychometric aspects. Specifically, 70% of the items met the validity criteria, while 30% were classified as invalid. Furthermore, the difficulty level analysis showed that most items were categorized as easy (65%), with no items reaching the difficult category. More importantly, the discrimination power analysis indicated that 75% of the items exhibited poor discriminatory ability, limiting the instrument's capacity to distinguish between students with different levels of achievement.

The study's most significant finding is that high reliability did not necessarily correspond to high overall test quality. While the instrument consistently measured student performance, many items failed to adequately discriminate between high- and low-achieving students and did not provide a balanced representation of item difficulty. This finding suggests that discrimination power and item difficulty distribution may have a greater influence on assessment quality than is often assumed in classroom-based testing practices. Consequently, the results challenge the common assumption that a reliable test instrument is automatically a high-quality instrument and highlight the importance of comprehensive psychometric evaluation in educational assessment.

From a scientific perspective, this study strengthens previous findings regarding the importance of item analysis in improving educational assessment quality. At the same time, it questions the adequacy of relying solely on reliability coefficients as indicators of test effectiveness. By providing empirical evidence from the context of Islamic Religious Education in vocational secondary education, this study contributes to the growing body of literature on educational measurement and reinforces the need to evaluate validity, reliability, difficulty level, and discrimination power simultaneously. The findings also enrich scholarly discussions concerning the application of Classical Test Theory in school-based assessment and the development of evidence-based evaluation practices in Islamic education.

Despite these contributions, the study has several limitations. First, the research involved a relatively small sample of 23 students from a single vocational school, which limits the generalizability of the findings to broader educational contexts. Second, the analysis focused exclusively on Classical Test Theory indicators and did not include distractor effectiveness analysis or more advanced psychometric approaches such as Item Response Theory (IRT). Therefore, future studies involving larger and more diverse samples, multiple educational settings, and additional psychometric analyses are necessary to provide a more comprehensive understanding of assessment quality in Islamic Religious Education.

Based on the findings of this study, Islamic Religious Education teachers are encouraged to revise invalid items, improve items with poor discrimination power, and develop a more balanced distribution of item difficulty levels to enhance assessment quality. Schools should support these efforts by providing professional development programs on test construction, educational measurement, and psychometric analysis, as well as by establishing systematic item review procedures and item banks. Furthermore, future researchers are encouraged to conduct studies with larger and more diverse samples, incorporate distractor effectiveness and essay-item analyses, and explore advanced psychometric approaches such as Item Response Theory (IRT) to obtain a more comprehensive understanding of assessment quality in Islamic Religious Education.

References

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Azizah, & Supahar. (2024). Analisis kualitas butir soal penilaian harian bersama fisika kelas X SMA Negeri 1 Patikraja. *Jurnal Pendidikan Fisika*, 10(2), 306–312.

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice, 25*(6), 551–575.

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education, 3*, Article 22.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications* (5th ed.). Sage.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hermanto, H., Anggraeni, D., & Sutirman, S. (2021). Analisis butir soal dengan classical test theory untuk mengukur kemampuan kognitif siswa. *Efisiensi: Kajian Ilmu Administrasi, 18*(1), 1–14.

Imam, M. K. (2023). Hasil uji tingkat daya pembeda, tingkat kesukaran, keberfungsian distraktor, validitas dan reliabilitas pada soal PKN siswa kelas IV di SDN 066056 Medan Denai. *Jurnal PGSD Indonesia, 9*(1), 42–46.

Jayusman, I., & Shavab, O. A. K. (2020). Studi deskriptif kuantitatif tentang aktivitas belajar mahasiswa dengan menggunakan media pembelajaran Edmodo dalam pembelajaran sejarah. *Jurnal Artefak, 7*(1), 13–20.

Lestari, M., Halini, H., & Indriani, T. (2023). Analisis tingkat kesukaran soal persamaan dan pertidaksamaan nilai mutlak melalui pendekatan teori tes klasik. *Indonesian Journal of Learning and Instructional Innovation, 1*(2), 28–33.

M., B. (2017). Evaluasi belajar peserta didik (siswa). *Jurnal Idaarah, 1*(2), 257–267.

Naga, D. S. (2004). Peranan interkorelasi butir terhadap koefisien reliabilitas Cronbach Alpha dan Kuder–Richardson. *Jurnal Kependidikan, 34*(2), 123–127.

Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of students* (7th ed.). Pearson.

Prabowo, I. (2024). Analisis kualitas soal asesmen sumatif pendidikan agama Islam dan budi pekerti. *Jurnal Ilmiah Insan Mulia, 1*(1), 1–7.

Putri, S., Zulyusri, Z., & Violita, V. (2023). Analisis kualitas soal ujian tengah semester biologi kelas XII SMA Negeri 6 Padang tahun pelajaran 2021/2022. *Jurnal Metaedukasi: Jurnal Ilmiah Pendidikan, 4*(1), 9–15.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3–13.

Rodriguez, M. C. (2016). Selected-response item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed.). Routledge.

Saputri, S., Larasati, L., & Shaleh, M. (2023). Analisis instrumen asesmen: Validitas, reliabilitas, tingkat kesukaran dan daya pembeda. *Didaktik: Jurnal Ilmiah PGSD FKIP Universitas Mandiri, 9*(5), 2986–2995.

Solichin, M. (2017). Evaluasi pembelajaran dalam pendidikan Islam. *Dirasat: Jurnal Manajemen dan Pendidikan Islam, 2*(2), 192–213.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education, 9*, Article 40.

Tosuncuoglu, I. (2018). Importance of assessment in ELT. *Journal of Education and Training Studies, 6*(9), 163–167.

Wahyuni, S., & Putri, N. D. (2020). Karakteristik instrumen penilaian akhir semester mata pelajaran matematika di SMKN 1 Braja Selebah. *LINEAR: Journal of Mathematics Education, 1*(2), 126–135.

Yulianti, N., Kosasih, E., & Hamdu, G. (2018). Analisis proses pembelajaran model latihan penelitian di sekolah dasar. *Pedadidaktika: Jurnal Ilmiah Pendidikan Guru Sekolah Dasar, 5*(2), 182–192.

Yusup, F. (2017). Uji validitas dan reliabilitas instrumen penelitian kuantitatif. *Jurnal Tarbiyah: Jurnal Ilmiah Kependidikan, 7*(1), 17–23.