

Analysis of the Quality of Final Semester Examination Items in Islamic Religious Education for Grade Vii at Smp N 4 Satu Atap Panai Hulu

Erni Agustin, Charles Rangkuti, Hadi Saputra Pangabea

Abstract

This study aims to analyze the quality of the Final Semester Examination (UAS) items for the Islamic Religious Education (PAI) subject in Grade VII at SMP N 4 Satu Atap Panai Hulu based on validity, reliability, item difficulty, discrimination power, and distractor effectiveness. This study employed a quantitative approach using a descriptive-evaluative method. The research data were obtained through documentation of the examination items, test blueprint, answer key, and students' answer sheets, as well as interviews with the PAI teacher as supporting data. The analysis was conducted through a logical approach covering the aspects of content, construction, and language, as well as an empirical analysis based on students' test results. The findings show that, in general, the Grade VII PAI Final Semester Examination instrument is categorized as reasonably suitable for use as an instrument for assessing learning outcomes. In terms of content, construction, and language, most of the test items have met the basic criteria for good test-item construction. However, the empirical analysis indicates that there are still weaknesses in empirical validity, reliability, item difficulty, discrimination power, and distractor effectiveness. The majority of the items fall into the easy to moderate difficulty categories, while relatively few items are classified as difficult. The dominance of items measuring lower cognitive levels also results in low discrimination power for several items and makes the instrument less effective in distinguishing students' abilities. Therefore, improvement and development of the test items are necessary by considering a balance of cognitive levels, item difficulty, discrimination power, and distractor effectiveness so that the PAI assessment instrument becomes more valid, reliable, objective, and comprehensive.

Keywords: Item Analysis, Item Quality, Final Semester Examination, Islamic Religious Education.

Erni Agustin ¹

¹ Magister Pendidikan Agama Islam, Universitas Pembangunan Panca Budi, Indonesia
e-mail: muhammadiqram628@gmail.com¹

Charles Rangkuti ² Hadi Saputra Pangabea³

^{2,3} Master of Management, Universitas Pembangunan Panca Budi, Indonesia
e-mail: charles@dosen.pancabudi.ac.id², hadi@dosen.pancabudi.ac.id³

3rd International Conference on the Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Navigating The Future: Business and Social Paradigms in a Transformative Era.
<https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF>

INTRODUCTION

Islamic Religious Education (PAI) at the junior high school level plays a fundamental role in shaping students' character, morality, and spiritual competence (Gusnanda, 2025). Religious education is not merely a process of transferring knowledge about religious teachings, but also a process of internalizing ethical, moral, and religious-cultural values that serve as guidelines for students' lives (Samsuddin, 2025). In this context, learning evaluation serves as a key instrument for determining the extent to which educational objectives have been achieved. The end-of-semester examination (UAS) is one of the strategic forms of assessment because it serves as a benchmark for measuring the success of learning throughout the semester (Nugroho, 2025). However, the effectiveness of the end-of-semester examination largely depends on the quality of the test items, including their validity, reliability, level of difficulty, discriminating power, and distractor effectiveness.

The learning evaluation system plays an important role in determining the effectiveness of the learning process, particularly in assessing students' learning outcomes. A well-designed evaluation should not only measure students' achievement of competencies but also employ high-quality test instruments to ensure that the assessment results are valid and reliable. Therefore, analysis of test-item quality is an important evaluative process for determining the extent to which the questions used in the end-of-semester examination can accurately and fairly measure students' competencies.

Education aims to develop students' potential comprehensively; therefore, every component of an evaluation instrument should reflect fairness, equality, and accuracy in measuring learning outcomes (Reza Alinata et al., 2024). In the context of Islamic Religious Education (PAI), the quality of test items should be able to assess not only students' cognitive abilities but also their understanding of religious values, which constitute an essential objective of the subject. Evaluation instruments that lack validity or reliability may result in inaccurate interpretations of students' learning outcomes, causing examination results to fail to represent their actual abilities. Previous research has indicated that the quality of PAI test items often needs to be improved in terms of validity and reliability in order to produce meaningful assessment results (Cahyaningrum, 2019).

This study is based on national regulations governing educational assessment standards. The Regulation of the Minister of Education and Culture Number 23 of 2016 concerning Educational Assessment Standards stipulates that assessment should be valid, objective, transparent, and accountable (UUD RI, 2016). In this context, teachers and schools have a legal responsibility to ensure that every assessment instrument undergoes adequate item-quality analysis before being used to determine students' final grades. This requirement highlights the importance of evidence-based assessment and scientifically sound procedures, ensuring that PAI assessment does not merely serve as an administrative formality but also functions as a strategic instrument for character development and the enhancement of students' competencies (Permendikbud, 2016).

The problem underlying this study arises from actual practices in the field. Based on initial observations and interviews, PAI teachers at SMP N 4 Satu Atap Panai Hulu frequently develop end-of-semester examination questions without conducting adequate item-analysis procedures, such as validity testing, reliability testing, analysis of item difficulty, discriminating power, and distractor effectiveness. Consequently, some questions were found to be excessively easy or predominantly focused on lower-order cognitive processes (C1—recall), while some distractors were ineffective and several items were unable to distinguish between high- and low-achieving students. This finding is consistent with previous research

indicating that many schools have not yet optimized the procedures for analyzing PAI test items, which may negatively affect the overall quality of assessment (Alacrity, 2024).

This study aims to examine the quality of the examination questions based on five main aspects of evaluation-instrument analysis: content validity, reliability, level of difficulty, discriminating power, and distractor effectiveness. The analysis is expected to provide useful feedback for teachers and school administrators in improving the quality of test items, while also supporting the development of a higher-quality question bank in the future. The findings are also expected to contribute to the improvement of the overall quality of Islamic Religious Education learning at SMP N 4 Satu Atap Panai Hulu.

This mini research emphasizes the importance of test-item quality analysis as part of efforts to improve the quality of Islamic Religious Education at SMP N 4 Satu Atap Panai Hulu. Through systematic item analysis, teachers can identify the strengths and weaknesses of the questions used, thereby ensuring that the assessment process is more valid, reliable, and fair. This study is expected to make a meaningful contribution to improving the quality of PAI learning, supporting teachers' professional development, and fostering students' character so that they possess noble moral values and the competencies necessary to face the challenges of the times. These efforts are in line with the objectives of national education, namely to produce a young generation that is intellectually capable, possesses strong character, and demonstrates religious values.

LITERATURE REVIEW

2.1 Evaluation of Learning Outcomes

a. Definition of Evaluation

The term *evaluation* comes from the English word *evaluation*, which refers to assessment, appraisal, or the process of determining the value of a particular object (Rahman, 2019). In the context of education, evaluation has a broader and more profound meaning than merely assigning grades to students. Evaluation is a conscious, planned, systematic, and continuous process of collecting information about the learning process and its outcomes, which is subsequently used as a basis for making educational decisions. Therefore, evaluation cannot be separated from the overall learning process, because the success of learning can only be determined through appropriate evaluation activities.

Educational evaluation can be defined as an activity designed to determine the extent to which learning objectives have been achieved in accordance with previously established plans. In its implementation, evaluation does not only assess the final outcomes of learning but also examines the learning process taking place in the classroom (Siyami & Wally, 2024). Through evaluation, educators can determine whether the learning strategies employed have been effective, whether the instructional materials have been understood by students, and whether the learning objectives have been achieved optimally. Thus, evaluation serves as an instrument for continuously improving the quality of learning.

Evaluation also plays an important role as a means of quality control in education. Through evaluation results, schools and educators can assess the quality of the learning process that has been implemented. In addition, evaluation serves as a basis for determining educational policies at the classroom, school, and broader institutional levels. Therefore, the evaluation of learning outcomes is a process that encompasses measurement, assessment, and interpretation of students' abilities after they have participated in learning activities over a certain period. The results of this process should be academically and pedagogically accountable.

2.2 Objectives and Functions of Learning Outcomes Evaluation

The primary objective of learning outcomes evaluation is to obtain accurate, objective, and reliable information regarding the level of students' competency achievement. Such

information is essential because it serves as a basis for educators to assess the success of the learning process that has been implemented. Without evaluation, teachers cannot determine with certainty whether the learning objectives have been achieved or identify which aspects of the learning process need to be improved or enhanced (Warsah, 2022).

In general, the evaluation of learning outcomes aims to determine the extent to which students have mastered the learning materials provided, including knowledge, attitudes, and skills. In addition, evaluation aims to assess the effectiveness of the methods, strategies, and learning media employed by teachers during the learning process. The results of evaluation are used as a basis for determining appropriate follow-up actions, such as remedial programs for students who have not achieved the required level of mastery or enrichment programs for students who have achieved optimal learning outcomes (Rohman, 2020). Evaluation also functions as a form of academic accountability to relevant stakeholders, including schools, parents, and the wider community.

The functions of learning outcomes evaluation can be viewed from various perspectives. For teachers, evaluation serves as a tool for determining the level of success of their teaching and identifying weaknesses and shortcomings in the learning process. For students, evaluation functions as feedback that enables them to understand their own abilities and motivates them to improve their learning outcomes. For schools, evaluation serves as an indicator of educational quality and as a basis for planning and developing school programs. Meanwhile, for parents and the community, evaluation results provide information regarding the quality of education provided by the school.

2.3 Principles of Learning Outcomes Evaluation

To ensure that the evaluation of learning outcomes is conducted effectively and produces objective information, its implementation must be guided by established evaluation principles. These principles serve as a foundation to ensure that evaluation is not merely a formal requirement but genuinely contributes to improving the quality of learning. The principles that should be considered include continuity, objectivity, comprehensiveness, fairness, transparency, and accountability.

The principle of continuity emphasizes that evaluation should not only be conducted at the end of the learning process but continuously throughout the learning process (Idrus, 2019). Through continuous evaluation, teachers can monitor students' learning progress more accurately. The principle of objectivity requires evaluation to be based on actual data and facts rather than personal feelings or the evaluator's subjectivity. The principle of comprehensiveness requires evaluation to cover all aspects of students' abilities, including the cognitive, affective, and psychomotor domains.

In addition, evaluation must be conducted fairly, meaning that it should neither disadvantage nor unfairly benefit particular students. The principle of transparency requires evaluation results to be accessible to and understandable by students, enabling them to recognize their learning position and progress. The principle of accountability requires evaluation results to be scientifically, administratively, and ethically defensible, thereby ensuring that evaluation functions as a credible instrument in the educational process.

B. Tests as Instruments for Evaluating Learning Outcomes

A test is one of the most frequently and widely used evaluation instruments in the learning process at various levels of education, ranging from primary education to higher education (Armedi, 2025). The use of tests is considered effective because they can provide a measurable picture of students' learning outcomes within a relatively short period. A test functions as a means of measuring students' learning outcomes through a series of questions, statements, or tasks that are systematically and deliberately constructed based on learning objectives and predetermined competencies. Therefore, tests should not be developed

arbitrarily but should be designed according to clear and well-defined learning achievement indicators.

Through testing, teachers can obtain objective information regarding students' levels of knowledge, conceptual understanding, and mastery of the learning materials taught during a particular period. This information serves as a basis for determining the extent to which learning objectives have been achieved and identifying areas of the material that still require reinforcement. In addition, test results can be used to identify differences in students' abilities, enabling teachers to design follow-up learning activities that are more appropriate to students' needs and characteristics.

In addition to functioning as a tool for measuring learning achievement, tests play a strategic role in the overall learning process by serving as a means of feedback for both teachers and students (Christine & Sitanggang, 2025). For teachers, test results can be used to evaluate the effectiveness of teaching methods, the suitability of the instructional materials delivered, and the teaching strategies implemented in the classroom. If test results indicate that students' achievement has not reached the expected level, teachers can make improvements through remedial programs, enrichment activities, or adjustments to their instructional approaches. For students, tests can serve as a means of self-reflection, allowing them to identify their strengths and weaknesses in understanding the learning materials. This can encourage learning motivation and awareness to improve academic achievement continuously.

As an evaluation instrument used in making various educational decisions, a test must meet several important requirements to ensure that the results obtained are trustworthy, accurate, and scientifically accountable. These requirements include validity, reliability, objectivity, and practicality. A test with high validity is capable of measuring the competencies that it is intended to measure in accordance with the learning objectives and predetermined indicators. A reliable test produces consistent scores when administered to groups of students or under relatively similar conditions. An objective test is not influenced by the evaluator's subjectivity, thereby making the assessment results fairer and more transparent. Meanwhile, a practical test is easy to develop, administer, and score without requiring excessive time, effort, or financial resources. When these requirements are fulfilled, a high-quality test can provide an accurate, comprehensive, and overall picture of students' learning outcomes.

C. Test-Item Analysis

Test-item analysis is a systematic and planned activity undertaken to examine and evaluate each test item used in an assessment in order to determine the empirical quality of individual items (Afrilianti, 2025). This activity aims to determine whether a particular test item meets the criteria for a good evaluation instrument, is appropriate for use, and is consistent with the learning objectives, or whether it needs to be revised or even eliminated because it fails to meet the expected quality standards. Therefore, test-item analysis is an important step in ensuring the quality of learning assessment.

Through test-item analysis, teachers and researchers can ensure that the questions used are genuinely capable of measuring students' abilities accurately, fairly, and proportionally. This analysis also plays an important role in improving the overall quality of learning assessment. High-quality test items produce accurate and valid data regarding students' abilities, allowing educational decisions based on test results, such as determining graduation, remedial programs, or enrichment activities, to be academically justified. Conversely, test items that have not undergone an analysis process may result in measurement errors that can disadvantage students and reduce the credibility of the learning evaluation process.

The aspects analyzed in test items include validity, reliability, level of difficulty, discriminating power, and distractor effectiveness (Magdalena, 2001). Validity indicates the extent to which a test item is capable of measuring the competency targeted by the learning objectives. Reliability indicates the consistency of the overall test results when the test is

administered under relatively similar conditions. The level of difficulty describes the proportion of students who are able to answer an item correctly, thereby indicating whether the item is classified as easy, moderate, or difficult. Discriminating power indicates the ability of an item to distinguish between students with high and low levels of ability. Meanwhile, distractor effectiveness indicates the extent to which incorrect answer options function effectively as distractors for students who have not yet mastered the learning material, thereby maintaining the quality of multiple-choice test items.

D. Islamic Religious Education

Islamic Religious Education (PAI) is a subject that plays a strategic and fundamental role in the national education system, particularly in developing students' faith, piety, and noble character (Aris, 2022). This subject is designed to instill Islamic teachings and values comprehensively so that students do not merely understand religion as a body of knowledge but also recognize it as a guide for life reflected in their attitudes and daily behavior. Through Islamic Religious Education, students are expected to develop spiritual, moral, and social awareness in accordance with Islamic values.

The primary objective of Islamic Religious Education is to guide students toward developing a correct, comprehensive, and in-depth understanding of Islamic teachings, including aspects of *aqidah* (faith), worship, morality, and *muamalah* (social and interpersonal relations). This understanding is expected not to remain at the conceptual level but to be internalized and manifested in students' daily lives (Hapsari, 2025). Therefore, PAI learning is not solely oriented toward the mastery of cognitive aspects but also encompasses the development of affective and behavioral dimensions, allowing Islamic values to become fully embedded in students' personalities.

In line with these objectives, evaluation in Islamic Religious Education must be designed accurately, systematically, and comprehensively so that it can measure the achievement of learning objectives in a balanced and holistic manner. PAI evaluation is not only intended to assess students' level of knowledge but also to determine the extent to which Islamic teachings have been understood, internalized, and applied in real-life situations (Faridah & Rizqi, 2025). Therefore, the forms and instruments of PAI assessment must be designed by taking into account the characteristics of the subject matter and the holistic nature of its learning objectives.

Islamic Religious Education test items are required to meet high-quality standards so that the assessment results genuinely reflect students' abilities objectively and accurately. High-quality test items should be capable of measuring students' ability to understand Islamic teachings, internalize their values, and apply them in their attitudes and everyday behavior. Thus, a well-designed and properly implemented evaluation process functions not only as a tool for measuring learning outcomes but also as a means of developing and strengthening students' religious character, enabling the objectives of Islamic Religious Education to be achieved optimally.

RESEARCH METHODOLOGY

This study employs a quantitative approach using a descriptive-evaluative method. A quantitative approach is used because this study analyzes data in the form of students' responses to the End-of-Semester Examination (UAS) in Islamic Religious Education (PAI) for Grade VII students at SMP N 4 Satu Atap Panai Hulu. The analysis is conducted to determine the quality of the test items based on validity, reliability, level of difficulty, discriminating power, and distractor effectiveness. The descriptive-evaluative method is used to describe and evaluate the quality of the test items based on empirical data obtained from students' responses. In addition to quantitative data, this study also employs interviews with the PAI teacher as

supporting data to obtain information regarding the process of developing the test items and the implementation of learning assessment.

Based on the problems identified in the development and use of the End-of-Semester Examination questions for Islamic Religious Education in Grade VII at SMP N 4 Satu Atap Panai Hulu, this study proposes the hypothesis that the quality of the test items used does not yet fully meet the criteria for a good evaluation instrument. This assumption is based on the indication that the development of the test items was not preceded by systematic item analysis, meaning that the quality of the questions may not yet adequately reflect students' abilities objectively and accurately (Sugiyono, 2020). In terms of validity and reliability, it is suspected that some test items still have weaknesses, either because they are not fully aligned with the indicators and competencies being measured or because the instrument as a whole has not demonstrated an optimal level of consistency.

Furthermore, the level of difficulty of the test items is expected to be disproportionate because some questions may be too easy or too difficult. Such conditions may prevent the test from measuring students' abilities in a balanced manner. This study also hypothesizes that the discriminating power and distractor effectiveness of the multiple-choice questions in the Grade VII PAI End-of-Semester Examination have not functioned optimally. Some test items may still be unable to distinguish between students with high and low levels of ability, while the distractor options may not function effectively because they are less capable of attracting students who have not yet mastered the learning material. Therefore, the overall quality of the test items is expected to require improvement so that the PAI learning assessment instrument can produce measurements that are more valid, reliable, objective, and fair.

This study uses documentation and testing as the primary data collection techniques, while interviews are employed as supporting data. Documentation is conducted by collecting various documents related to the implementation of the Grade VII PAI End-of-Semester Examination at SMP N 4 Satu Atap Panai Hulu, including the PAI examination paper, test blueprint, answer key, and students' answer sheets. These documents serve as the primary data sources for analyzing the quality of the test items, particularly in determining the alignment of the questions with learning indicators, the competencies being measured, and the subject matter assessed.

The testing technique utilizes students' responses to the Grade VII PAI End-of-Semester Examination (Naamy, 2019). Students' answer sheets are analyzed to obtain empirical data concerning the characteristics of each test item. The test results are then used to calculate and analyze validity, reliability, level of difficulty, discriminating power, and distractor effectiveness. By using students' actual test responses, test-item analysis can be conducted objectively based on real data obtained from the field.

To complement and strengthen the results of the analysis, the researcher also conducted an interview with the Islamic Religious Education teacher. The interview aimed to obtain information regarding the process of developing the test items, the teacher's considerations in determining the form and level of difficulty of the questions, and the challenges encountered in analyzing test-item quality. The interview data were subsequently used as supporting qualitative data to clarify and provide context for the findings of the test-item analysis. Thus, the results of the study are expected to provide a more comprehensive description of the quality of the Grade VII PAI End-of-Semester Examination questions at SMP N 4 Satu Atap Panai Hulu.

RESULTS AND DISCUSSION

4.1 Results of the Analysis of the Grade VII PAI End-of-Semester Examination Instrument

The quality analysis of the Grade VII Islamic Religious Education (PAI) End-of-Semester Examination (UAS) instrument at SMP N 4 Satu Atap Panai Hulu was conducted systematically and comprehensively using two main approaches, namely logical (rational) analysis and empirical analysis. The use of these two approaches was intended to ensure that the assessment of instrument quality was not merely normative and theoretical but was also supported by empirical evidence obtained from students' test results.

Logical or rational analysis constitutes the initial stage, which aims to assess the quality of test items based on academic and pedagogical considerations before relating them to statistical data (Arifin, 2021). This analysis covers three main domains: substance or content, construction, and language. At this stage, each test item is examined to ensure its alignment with learning indicators, basic competencies, PAI learning objectives, and the principles of good and appropriate test-item construction. Logical analysis is particularly important because even test items that appear statistically sound may lose their educational significance if their content, construction, or language fails to meet academic standards.

Meanwhile, empirical analysis is conducted to examine the quality of test items based on actual data obtained from students' responses, using statistical formulas commonly applied in test-item analysis (Nasir, 2022). Empirical analysis includes validity, reliability, level of difficulty, discriminating power, and distractor effectiveness. Through this analysis, it can be determined to what extent each test item is capable of measuring students' competencies accurately, consistently, and objectively, as well as its ability to distinguish students according to their levels of ability.

This dual approach was selected because the quality of an assessment instrument cannot be determined solely by the conformity of its content with the curriculum or textbooks. A good assessment instrument must be capable of producing information that is valid, reliable, and fair for all students. Therefore, the integration of logical and empirical analysis serves as an important foundation for assessing the feasibility of the Grade VII PAI End-of-Semester Examination instrument as a tool for evaluating students' learning outcomes.

4.2 Results of the Test-Item Analysis

a. Substance/Content Domain

The results of the analysis of the substance or content domain indicate that, in general, the Grade VII PAI End-of-Semester Examination items are consistent with the learning indicators and basic competencies stated in the curriculum. The materials assessed reflect the scope of PAI learning at the junior high school level, covering aspects of *aqidah* (faith), morality, *fiqh* (Islamic jurisprudence), and Islamic history. These four aspects constitute important components of PAI learning and serve as foundations for developing students' faith, piety, and noble character.

In terms of content alignment, the questions assess basic concepts that are expected to be mastered by Grade VII students, such as the recognition of *Asmaul Husna*, values of commendable character, procedures for purification and congregational prayer, as well as basic knowledge of the Prophet Muhammad (peace be upon him) and angels. In addition, none of the test items contain elements related to ethnicity, religion, race, and intergroup relations (SARA), pornography, violence, or other negative content. Thus, the items have generally fulfilled assessment principles based on ethics, fairness, and respect for educational values.

However, upon closer examination, it was found that the majority of the questions remain oriented toward measuring factual and conceptual knowledge, such as recalling definitions, identifying terms, or recognizing basic information. Questions requiring students to understand the meaning of Islamic values more reflectively, relate them to everyday life contexts, or demonstrate the moral implications of particular teachings remain relatively limited. This condition indicates that the function of assessment as a means of evaluating the internalization and application of Islamic values has not yet been fully optimized.

The implication of this finding is that the PAI End-of-Semester Examination instrument tends to assess lower-level cognitive aspects more heavily than the holistic objectives of PAI learning, which aim to develop students who not only know Islamic teachings but also understand, believe in, and apply them in their daily lives. Therefore, from the perspective of content, the instrument still requires enrichment so that it can measure PAI competencies more comprehensively.

4.2 Construction Domain

From the perspective of test-item construction, most of the questions have fulfilled the principles of good test-item writing. The stems are generally formulated concisely, clearly, and directly address the problem being asked, thereby minimizing ambiguity for students. Furthermore, the test items do not depend on one another, meaning that each question stands independently and is not influenced by the answer to a previous question.

The multiple-choice questions also avoid the use of double-negative statements that could potentially confuse students. The lengths of the answer options are relatively similar, and each item has only one correct answer. This indicates that, technically, the test construction generally follows the guidelines for developing assessment instruments commonly used in educational assessment.

Nevertheless, the analysis also identified several weaknesses in test-item construction, particularly in questions that are overly simple and direct. Such questions tend to provide implicit clues to the correct answer, either through the wording of the question stem or through answer options that are too obviously contrasting. This type of construction weakness affects distractor effectiveness because students can easily eliminate the incorrect options.

In addition, questions that are too direct may affect their discriminating power because both high- and low-ability students have relatively equal opportunities to answer them correctly. Therefore, although the technical construction of the questions is generally adequate, further refinement is necessary to enable the test items to challenge students' thinking abilities more effectively.

4.3 Language Domain

From the linguistic perspective, the Grade VII PAI End-of-Semester Examination instrument generally uses standard and communicative Indonesian that is appropriate for the students' level of development. The sentence structures used in the question stems and answer options are relatively simple and easy to understand, thereby minimizing the possibility of misinterpretation or misunderstanding of the questions.

The use of religious terminology is also appropriate and consistent with the PAI materials taught in class. No regional language, ambiguous expressions, or words that could potentially offend students were identified. This indicates that, from the linguistic perspective, the instrument has generally fulfilled the principles of readability and clarity.

However, several test items still contain repeated words or phrases in the answer options, which may reduce the quality of the item construction. Such repetition can provide indirect clues to students regarding the correct answer. Therefore, although the language aspect is generally good, refinement of the wording remains necessary to further optimize the quality of the assessment instrument.

4.4 Results of the Empirical Analysis of Test Items

Empirical analysis was conducted to objectively determine the quality of the Grade VII Islamic Religious Education (PAI) End-of-Semester Examination (UAS) items based on students' test results (Nurul, 2021). This analysis employs a quantitative approach using statistical formulas commonly applied in test-item analysis, including validity, reliability, level of difficulty, discriminating power, and distractor effectiveness. The empirical approach is important because it provides an actual picture of the performance of each test item in

measuring students' abilities while complementing the theoretical findings obtained through logical analysis.

4.5 Test-Item Validity

The validity of the multiple-choice test items was analyzed using the point-biserial correlation formula, which is formulated as follows:

This formula is used to determine the extent to which the score obtained on an individual test item correlates with the total test score. A high correlation coefficient indicates that the test item is capable of measuring an ability that is consistent with the overall objective of the test. In other words, the higher the value of r_{bis} , the better the quality of the test item in distinguishing students according to their level of mastery of the learning material.

The results of the analysis indicate that most of the test items have validity coefficients within the moderate category, particularly those measuring basic knowledge of Islamic Religious Education. Empirically valid test items demonstrate a positive relationship between the item score and the total test score, indicating that students who obtain high overall test scores tend to answer these items correctly. This finding suggests that, in general, the instrument is capable of measuring competencies that are aligned with the objectives of PAI learning assessment.

However, several test items were also found to have low validity coefficients. The low r_{bis} values of certain items were caused by the characteristics of the questions, which were too easy and did not sufficiently require students to engage in cognitive processing. Such questions tend to be answered correctly by almost all students, regardless of whether they have high or low ability levels. Consequently, the item scores do not correlate significantly with the total test scores. This condition demonstrates that although a test item may possess content validity, it is not necessarily empirically valid if it fails to reflect differences in students' abilities.

For essay questions, validity was analyzed using the *product-moment correlation* formula, both in deviation-score and raw-score forms. The results indicate that the validity of essay questions is strongly influenced by the clarity of the expected answer boundaries, their alignment with learning indicators, and consistency in the scoring process. Essay questions with clear indicators and structured scoring guidelines tend to demonstrate better validity than questions that are overly general or open-ended without clear limitations.

4.6 Reliability of the Test Instrument

The reliability of the multiple-choice test instrument was analyzed using the Kuder-Richardson 20 (KR-20) formula, which is formulated as follows:

$$r_{11} = \left(\frac{n}{n-1} \right) \left(\frac{S^2 - \sum p_i q_i}{S^2} \right)$$

The KR-20 formula is used to measure the internal consistency of a test, namely the extent to which the items within a test consistently measure the same construct. The calculation results indicate that the reliability of the Grade VII PAI End-of-Semester Examination instrument is generally within the moderate category, indicating that the test has a relatively adequate level of consistency as an instrument for measuring learning outcomes.

Nevertheless, the reliability of the instrument has not yet reached the high category. One factor contributing to this condition is the predominance of test items with a low level of difficulty, meaning that many of the questions are relatively easy. Questions that are too easy tend to produce relatively homogeneous student scores, resulting in limited variation in scores among students. Low score variability can lead to a decrease in the reliability coefficient

because the instrument becomes less sensitive to consistently detecting differences in students' abilities.

For essay questions, reliability was analyzed using Cronbach's Alpha formula. The results indicate that the reliability of essay questions is highly dependent on the availability and quality of scoring rubrics or scoring guidelines. Insufficiently detailed scoring guidelines for some questions may increase the potential for scorer subjectivity, thereby reducing the consistency of the assessment results. This finding indicates that improving the reliability of essay questions depends not only on the quality of the questions themselves but also on the clarity and accuracy of the scoring procedures.

4.7 Item Difficulty

The difficulty level of the test items was analyzed using the following formula:

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

This formula is used to determine the proportion of students who are able to answer a particular test item correctly. A high P value indicates that the item is relatively easy, whereas a low P value indicates that the item is relatively difficult.

The results of the analysis indicate that the majority of the test items fall into the easy category ($P > 0.70$) to moderate category ($0.31 \leq P \leq 0.70$). Only a very small number of items are classified as difficult. The predominance of easy items indicates that the PAI End-of-Semester Examination instrument tends to place greater emphasis on mastery of basic materials and memorization, while providing insufficient challenges for students with higher academic abilities.

An unbalanced distribution of item difficulty directly affects the quality of the instrument as a measurement tool. A test dominated by easy questions becomes less sensitive in representing the overall distribution of students' abilities. Consequently, the test results are less effective in providing useful diagnostic information for teachers when designing follow-up learning activities, such as enrichment and remedial programs.

4.8 Discriminating Power of Test Items

The discriminating power of the test items was analyzed using the following formula:

$$D = PA - PB$$

This formula is used to determine the ability of a test item to distinguish between students in the upper and lower groups based on their test scores. A high discriminating power value indicates that the item is effective in distinguishing between high- and low-ability students.

The results of the analysis indicate that most of the test items have low to moderate discriminating power. Items with low discriminating power are generally questions with a low level of difficulty, which are answered correctly by almost all students. This condition results in only a small difference in the proportion of correct responses between the upper and lower groups.

Conversely, several test items that relate concepts to examples of behavior or the application of Islamic values in everyday life demonstrate relatively better discriminating power. This finding indicates that questions requiring conceptual understanding and application tend to be more effective in distinguishing students' ability levels than questions that merely assess memorization.

The generally low level of discriminating power indicates that the selective function of the PAI End-of-Semester Examination instrument remains limited. The instrument is not yet fully capable of accurately identifying differences in students' abilities. Therefore, the quality of the learning assessment needs to be improved through better test-item design.

4.9 Distractor Effectiveness

Distractor effectiveness was analyzed by examining the proportion of students selecting each answer alternative in the multiple-choice questions. A distractor is considered effective when it is capable of attracting students who have not yet mastered the learning material and is selected by a proportion of students, particularly those in the lower-ability group.

The results of the analysis indicate that some distractors have not functioned optimally because they were rarely or never selected by students. Ineffective distractors are generally caused by concepts that are too far removed from the correct answer, the use of irrelevant terminology, or wording that is not sufficiently plausible. Such distractors can be easily eliminated by students, thereby reducing the quality of the test items.

Nevertheless, several questions were found to have effective distractors, particularly items in which the answer options contain conceptually similar alternatives. Well-designed distractors can mislead students who have not developed an in-depth understanding of the learning material, thereby improving the discriminating power and empirical validity of the test items.

4.10 Discussion of Research Findings

Based on the logical and empirical analyses conducted using a rational approach and relevant statistical formulas, it can be concluded that the quality of the Grade VII Islamic Religious Education (PAI) End-of-Semester Examination (UAS) instrument at SMP N 4 Satu Atap Panai Hulu is in the moderately adequate category for use, but it does not yet fully meet the standards of an ideal assessment instrument. These findings indicate that the instrument is capable of measuring some aspects of students' competencies but still has several limitations that require serious attention in order to improve the quality of learning assessment.

The results of the logical analysis indicate that, in terms of substance, construction, and language, the PAI End-of-Semester Examination questions have generally fulfilled the basic criteria for good test-item construction. The assessed materials are relevant to the curriculum and learning indicators, the language used is communicative and easy to understand, and the construction of the questions is relatively clear and does not create significant ambiguity. However, the findings from the empirical analysis show that the technical quality of the test items is not yet fully optimal, particularly in terms of empirical validity, discriminating power, and the balance of item difficulty.

The predominance of questions targeting lower-level cognitive processes is a major finding that affects the overall quality of the instrument. Questions that primarily measure students' ability to remember and understand cause many items to have low levels of difficulty and low discriminating power. Consequently, the instrument becomes less sensitive in distinguishing students according to their academic ability levels. This condition is reflected in the limited variation in students' scores and the instrument's limited ability to accurately identify high- and low-ability groups.

Furthermore, the imbalance in item difficulty has a direct impact on the construct validity of the instrument. An assessment instrument should represent the ability construct being measured in a proportional manner. However, the predominance of easy questions results in a relatively narrow construct being measured and fails to adequately represent the overall competencies expected from Islamic Religious Education learning. Therefore, although several test items have demonstrated content validity, the overall construct validity of the instrument still needs to be improved.

The findings of this study also emphasize that statistical test-item analysis plays a very important role in the learning assessment process. Statistical analysis not only serves to evaluate the quality of test items after they have been administered but also provides an objective basis for making decisions regarding the revision, replacement, or elimination of test items. Without empirical analysis, teachers may tend to evaluate the quality of questions

subjectively, making it difficult to systematically identify potential weaknesses in the assessment instrument.

In the context of Islamic Religious Education, these findings have broader implications. PAI learning is not only intended to improve students' knowledge but also to develop their attitudes, values, and Islamic behavior in everyday life. Therefore, PAI assessment instruments should be designed to measure higher-order thinking skills, understanding of religious values, and students' ability to apply Islamic teachings contextually. The limitations of the current instrument in measuring these aspects indicate the need for a shift in the approach to developing PAI assessment questions.

Furthermore, the findings regarding distractor effectiveness and the reliability of essay questions indicate that the technical aspects of test development still need to be improved. Ineffective distractors reduce the quality of multiple-choice questions, while insufficiently detailed scoring rubrics may introduce subjectivity into the assessment of essay questions. This demonstrates that improving the quality of an assessment instrument is not only related to the content of the questions but also to the accuracy of their design and scoring procedures.

Therefore, the findings of this study recommend that PAI teachers develop test items with a more balanced distribution of cognitive levels, pay greater attention to the quality of item construction and distractor effectiveness, and provide essay questions with clear and structured scoring rubrics. These efforts are expected to improve the validity, reliability, and meaningfulness of the assessment instrument so that the assessment results genuinely reflect the achievement of Islamic Religious Education objectives in a comprehensive and holistic manner.

CONCLUSION

Based on the results of the logical and empirical analyses of the Grade VII Islamic Religious Education (PAI) End-of-Semester Examination (UAS) instrument at SMP N 4 Satu Atap Panai Hulu, it can be concluded that, in general, the quality of the test items is in the moderately adequate category for use as an instrument for evaluating students' learning outcomes. In terms of substance, construction, and language, most of the test items are consistent with the learning indicators and the principles of good test-item construction. However, the results of the empirical analysis indicate that there are still limitations in terms of empirical validity, reliability, level of difficulty, discriminating power, and distractor effectiveness, particularly among items that are too easy and do not sufficiently require students to engage in deeper cognitive processes.

The predominance of questions targeting lower-level cognitive processes has contributed to low discriminating power and an imbalance in item difficulty, meaning that the instrument is not yet fully optimal in distinguishing students according to their academic abilities. In addition, suboptimal distractor effectiveness and insufficiently detailed scoring rubrics for essay questions also affect the overall quality of the instrument. Therefore, although the PAI End-of-Semester Examination instrument has met the minimum standards for assessment, continuous improvement is still necessary to ensure that the instrument can measure students' competencies in a more valid, reliable, and comprehensive manner in accordance with the objectives of Islamic Religious Education.

REFERENCES

- [1] Afrilianti, D. (2025). Karakteristik Tes Yang Baik. *Journal Of Innovative And Creativity*, 5(2), 1142–1158.
- [2] Arifin, Z. (2021). *Evaluasi Pembelajaran*.
- [3] Aris. (2022). *Ilmu Pendidikan Islam*. Penerbit Yayasan Wiyata Bestari Samasta.
- [4] Armadi. (2025). Karakteristik Tes Yang Baik Dan Proses Penyusunan Instrumen Tes

- Untuk Pembelajaran Di Sekolah. *PENDAGOGIA: Jurnal Pendidikan Dasar*, 5(April), 10–17.
- [5] Christine, I., & Sitanggang. (2025). Analisis Penggunaan Instrumen Test Sebagai Alat Evaluasi Pembelajaran Di Sekolah Dasar. *Jurnal Studi Multidisipliner*, 9(5), 65–72.
 - [6] Faridah, Z., & Rizqi, A. A. (2025). Konsep Moderasi, Integrasi Ilmu, Dan Kontekstualisasi Pendidikan Islam Dalam Pemikiran Muhammad Quraish Shihab. *Arsy*, 9(2), 93–110. <https://doi.org/10.32492/Arsy.V9i2.9202>
 - [7] Gusnanda, A. (2025). Implementasi Pembelajaran Pendidikan Agama Islam (PAI) Membentuk. *Jurnal Studi Islam Dan Humanior*, 6(1), 651–662. <https://doi.org/10.37680/Almikraj.V6i1.7783>
 - [8] Hapsari, T. T. (2025). Peran Pendidikan Agama Islam Dalam Membentuk Karakter Siswa Di Era Globalisasi. *Karakter : Jurnal Riset Ilmu Pendidikan Islam*, 2(3), 288–296. <https://doi.org/10.61132/Karakter.V2i3.1253>
 - [9] Idrus. (2019). Evaluasi Dalam Proses Pembelajaran Idrus L 1. *Jurnal Manajmenen Pendidikan Islam*, 2, 920–935.
 - [10] Magdalena, I. (2001). Analisis Instrumen Tes Sebagai Alat Evaluasi Pada Mata Pelajaran Sbdp Siswa Kelas Ii Sdn Duri Kosambi 06 Pagi. *Nusantara : Jurnal Pendidikan Dan Ilmu Sosial*, 3, 276–287.
 - [11] Naamy, N. (2019). Metode Penelitian Kualitatif Dasar-Dasar & Aplikasinya. In *Rake Sarasin* (Issue Maret). Pusat Penelitian Dan Publikasi Ilmiah LP2M. [https://repository.uinmataram.ac.id/2853/1/Buku Metode Penelitian.Pdf](https://repository.uinmataram.ac.id/2853/1/Buku%20Metode%20Penelitian.Pdf)
 - [12] Nasir, M. (2022). *Analisis Empirik Program Analisis Butir Soal Dalam Rangka Menghasilkan Soal Yang Baik Dan Bermutu Sebagai Alat Evaluasi Pembelajaran Fisika*. 336–347.
 - [13] Nugroho, A. (2025). Penggunaan Aplikasi Iteman Dalam Menganalisis Butir Soal Ulangan Akhir Semester Gasal TA 2023 / 2024 Ppkn Kelas X Di SMAN 1 Prambanan Sleman. *AGORA: Jurnal Kajian Pancasila Dan Kewarganegaraan Pendidikan*, 14(03), 810–822.
 - [14] Nurul, S. (2021). *Kesulitan Dan Daya Beda Butir Soal Ujian Akhir Semester*. 3, 198–214.
 - [15] Rahman, A. A. (2019). *Evaluasi Pembelajaran*. Uwais Inspirasi Indonesia.
 - [16] Reza Alinata, Winda Atika Sari, & Yuli Kartika Putri. (2024). Makna Pendidikan Dalam Perspektif Islam Dan Relevansinya Dengan Pendidikan Di Indonesia. *IHSANIKA : Jurnal Pendidikan Agama Islam*, 2(3), 169–182. <https://doi.org/10.59841/Ihsanika.V2i3.1416>
 - [17] Rohman, T. (2020). Konsep Evaluasi Program Pendidikan Islam. *Jurnal Literasiologi*, 3(3), 46–57.
 - [18] Samsuddin. (2025). Dasar-Dasar Pendidikan Islam Perspektif Hasan Langgulung Dan Relevansinya Di Era Disrupsi. *Jurnal Kajian Islam*, 2(1), 202–223.
 - [19] Siyami, F., & Wally, O. (2024). *Learning Evaluation Theory And Principles*. 7302–7306.
 - [20] Sugiyono. (2020). *Metodologi Penelitian Kuantitatif, Kualitatif Dan R & D*. Penerbit Alfabeta.
 - [21] UUD RI. (2016). *Peraturan Menteri Pendidikan Dan Kebudayaan Nomor 23 Tahun 2016 Tentang Standar Penilaian Pendidikan Menjelaskan Prinsip-Prinsip Penilaian Yang Harus Sahih, Objektif, Transparan, Dan Dapat Dipertanggungjawabkan*.
 - [22] Warsah, I. (2022). Evaluasi Pembelajaran (Konsep . Fungsi Dan Tujuan). *Jurnal Kajian Pendidikan Islam*, 1.