

Quality Assurance of Arabic School Examination Development: A Discrepancy Evaluation in Muhammadiyah Elementary Schools

Submitted 14 July 2026, Revised 30 July 2026, Accepted 1 August 2026

Putri Nabila Utami^{1*}, Supahar Supahar², Muhammad Salman Arrosyid³

^{1,2,3}Educational Research and Evaluation Program, Graduate School,
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
Corresponding Email: *putrinabila.2024@student.uny.ac.id

Abstract

The present study intended to evaluate the development of Grade VI Arabic School Examination items in Muhammadiyah elementary schools within the Special Region of Yogyakarta using the Discrepancy Evaluation Model. The focus of this evaluation was on assessing the degree of conformance to the set of criteria of the phases in developing test items, namely, design, installation, process, and product, and the discrepancies between what is done and what is supposed to be in order to generate recommendations on how to improve the quality assurance of the system of Arabic language assessment. The methodology of the study included conducting an evaluation research based on Malcolm Provus's Discrepancy Evaluation Model (DEM). The respondents in this research were Arabic test item writers, school principals, and the Vice Chair for the Al-Islam, Muhammadiyah Ideology, and Arabic (ISMUBA) Curriculum of the Primary and Nonformal Education Council of the Muhammadiyah Regional Board of the Special Region of Yogyakarta. They were selected purposively. Data collection methods were questionnaires, interviews, and reviewing documents. The content validity of the instruments was assessed using Content Validity Ratio (CVR) with the involvement of five expert validators, and all instruments were considered appropriate for use following their revision. Instrument reliability was calculated using Cronbach's Alpha. Quantitative data were analyzed descriptively on the basis of percentage achievement scores for every indicator of the evaluation, and qualitative data were analyzed by triangulating interview data and document review in order to detect discrepancies between the real condition and the standard for every stage of the evaluation. According to the results of the evaluation, the design phase conformed to all established standards, and there were no discrepancies between the real situation and the standards. Most components of the installation phase conformed to the set standards, but discrepancies appeared in relation to the readiness of resources in terms of the time devoted to the test development (57.14%), facilities (74.29%), and test development procedures understanding. As for the process phase, it was implemented in line with the set standards due to the alignment of test items with the test blueprint (93.99%) and the Standard Operating Procedures (76.79%), although there were still some shortcomings in regard to coordination, communication, and documentation of activities. At the product stage, most quality standards were achieved, including content appropriateness (95.77%), reviewer involvement (90.52%), and the completeness of archived test items and answer keys (100.00%). Nevertheless, technical documentation still requires improvement. Overall, the discrepancy analysis indicates that the implementation of Arabic School Examination item development at Muhammadiyah elementary schools in the Special Region of Yogyakarta has met the program standards and is feasible to continue, provided that improvements are made in resource management, coordination, documentation, and test quality control to further strengthen the quality assurance of Arabic language assessment.

Keywords: Discrepancy Evaluation Model, Test Item Development, Arabic School Examination, Assessment Evaluation.

INTRODUCTION

The quality of assessment is one of the main pillars of educational quality, as assessment results serve as a basis for decision-making at the student, teacher, school, and policymaker levels. From the perspective of educational measurement, high-quality assessments not only provide information on learning outcomes but also offer valid evidence to support the interpretation of scores and the appropriate use of assessment results (Ing et al., 2024; Otaya et al., 2020) . Thus, the quality of assessment cannot be understood merely as a

technical characteristic of an instrument, but as a system involving the planning, development, administration, interpretation, and utilization of assessment results. High-quality assessment practices must be supported by various sources of evidence that ensure the validity of score interpretation, ranging from content appropriateness, construct representation, and the development process to the utility of assessment results (Birenbaum, 2007; Sosa Paredes & Andersson, 2025) . This perspective broadens the concept of assessment quality from merely the quality of the product to the overall quality of the assessment process.

Developments in assessment research also indicate a paradigm shift from a measurement-oriented approach toward an approach that positions assessment as an integral part of enhancing learning and ensuring educational quality. Within this paradigm, validity and reliability remain important prerequisites but are no longer viewed as the sole indicators of assessment quality. Transparency of procedures, standardization, accountability, clarity of criteria, stakeholder involvement, and quality control mechanisms during the instrument development process are also dimensions that determine assessment quality (Appels et al., 2022) . Various studies indicate that assessment quality is a multidimensional construct influenced by the interaction between psychometric aspects, governance of administration, and the quality of the assessment process's implementation (Wang et al., 2025) . Therefore, ensuring assessment quality requires an evaluation that focuses not only on measurement outcomes but also on the overall mechanisms that produce the assessment instruments.

Although attention to assessment quality is increasing, research on instrument quality still largely focuses on the characteristics of assessment products, such as validity, reliability, difficulty level, discriminative power, distractor effectiveness, and the development of scoring rubrics (Patil et al., 2022; Rezigalla et al., 2024) . These approaches are important for ensuring that instruments possess adequate psychometric characteristics, but they do not fully explain how instrument quality is built through the development process. In fact, the quality of assessment instruments is also influenced by the quality of planning, the competence of the developers, team coordination, adherence to procedures, review and revision mechanisms, and documentation of the development process (Harris et al., 2024) . Thus, there is a need to shift the focus from evaluation that is solely oriented toward the assessment product to evaluation that also considers the assessment development process as part of the quality assurance system.

Quality assurance in test item development is a series of systematic procedures designed to ensure that each stage of instrument development produces assessments that are valid, fair, and aligned with the measurement objectives (Aithal & Aithal, 2020; Chambers et

al., 2026) . From the perspective of educational measurement, quality assurance is not only focused on examining the quality of the final product but also on quality control throughout the instrument development process—including the formulation of test specifications, item development , expert review, revision, pilot testing, documentation, and item bank management (Downing & Haladyna, 1997) . The implementation of such quality assurance requires quality control mechanisms at every stage of development, starting from establishing the assessment framework, drafting test specifications, item development, review, pilot testing, analysis of measurement results, revision, and finalization of the instrument (Phelps et al., 2020) . The development of high-quality assessment content also requires structured procedures, various forms of review and validation, and adequate documentation to maintain the consistency and credibility of the instrument (Twist & Fraillon, 2020) . Therefore, an evaluation of the item development process is necessary to determine whether instrument development standards are truly applied in practice.

This need becomes increasingly important in the administration of school exams. School exams are a form of summative assessment used to obtain information regarding students' achievement of competencies and to support decision-making at the school level (Babinčáková et al., 2020) . As an assessment with implications for educational decisions, the quality of school exams is determined not only by the statistical characteristics of the instrument but also by the rigor of the development process, which provides evidence of validity for the interpretation and use of assessment scores (Velasco, 2023) . Assessment validation requires various sources of evidence to support the interpretation and use of measurement results; therefore, the validation process must be an integral part of assessment development (Pepper, 2020) . Therefore, the administration of school exams requires a quality assurance system that ensures that assessment design, content development, item development, review, refinement, and documentation are carried out systematically and in a standardized manner (Lucander & Christersson, 2020) .

A number of previous studies have provided evidence of various issues regarding the quality of Arabic language assessments and assessments within the Muhammadiyah educational system. Lismawati(2019) , explained that through an analysis of ISMUBA test items at SD Muhammadiyah 5 in South Jakarta, she found variations in the difficulty level and discriminative power of the items. These findings indicate that the quality of the assessment instruments still requires attention so that the questions can provide accurate information regarding students' competency achievement. At a different educational level, Fikriyah (2021) found that the quality of the items in the Arabic Midterm Exam at

Muhammadiyah Junior High School 1 in Yogyakarta still needs to be analyzed in terms of validity and reliability. Meanwhile, Wasito (2019) indicates that issues in Arabic language learning assessment are not only related to the characteristics of test items but also to the implementation of the learning assessment system. These findings demonstrate that studies on Arabic language assessment within the Muhammadiyah context have focused on the quality of assessment instruments and the implementation of evaluations.

Research on Arabic language assessment also highlights the importance of aligning the competencies to be measured with the characteristics of the assessment instruments. At the elementary education level, Arabic language assessment covers various aspects of competency, such as vocabulary mastery (*mufradat*), linguistic structure (*qawa'id*), and text comprehension (*qira'ah*) (Robbany et al., 2025). The complexity of these constructs requires a systematic instrument development process to ensure that each test item represents the competencies that are the learning objectives (Qalbi et al., 2026). In the context of Muhammadiyah education, Arabic is part of the Al-Islam, Muhammadiyah Ideology, and Arabic (ISMUBA) curriculum; therefore, its development must take into account the interrelationship between learning objectives, language competencies, content, and student characteristics. Thus, the quality of Arabic language assessment is not determined solely by the quality of individual items but also by the overall alignment of the instrument development process with curriculum standards and assessment objectives.

Although these studies have made important contributions, there are limitations that need to be noted. First, the studies by Lismawati (2019) and Fikriyah (2021) primarily focused on the quality of the test item products through an analysis of item characteristics; thus, they did not explain how the test item development process was carried out prior to producing those items. Secondly, in Wasito's (2019) research the implementation of the evaluation system was considered, though the evaluation of the consistency of the item development standards and the processes of instrument development at each phase was not included. Thirdly, despite the large number of studies about the assessment of the Arabic language in the Muhammadiyah context, the aspect of the assessment process which is discussed is mostly limited to the discussion of assessment instruments or implementation of evaluations, whereas such aspects as the test item development process governance, resource readiness, coordination of the developers, the review and revision processes, procedural compliance, documentation and consistency of the developed test item with the standards remain under-researched. Therefore, there is a research gap in terms of the lack of evaluative

studies on test item development as a process in need of systematic comparison between the established standards and the real practice.

This research gap is even more relevant to the context of the development of the test items for the Arabic Language School Exam in Muhammadiyah elementary schools of the Special Region of Yogyakarta. The test item development process itself in the mentioned context takes place within the framework of the coordination process in educational units with ISMUBA subject teachers serving as developers and Muhammadiyah education management agency being one of the means for the standardization of the assessment quality process. There are many aspects involved in the process including planning, preparation by the developers, test blueprint construction, item writing, review and revision of the test items up to their finalization and documentation. Gaps may occur, for example, in the availability of resources, task distribution, developers' competencies, time allocation, coordination, the review process, adherence to procedures, or the completeness of documentation. However, research specifically mapping these gaps in the development process of the Arabic Language School Exam for Muhammadiyah Elementary Schools in the Special Region of Yogyakarta remains limited.

To address this gap, an evaluation approach is needed that not only assesses the quality of the final product but is also capable of identifying the alignment between standards and actual conditions at various stages of program implementation. One relevant model is the Discrepancy Evaluation Model developed by Malcolm Provus. This model views evaluation as a systematic process for comparing established standards (standard) with the actual conditions of program implementation (performance), thereby identifying discrepancies (discrepancy) that require corrective action (Mustafa, 2021). These characteristics align with the research needs because the test item development process can be viewed as a program with standards at the design, implementation, process, and product stages. Through comparison of standards and the situation in real practice at each step, the evaluation will be able to identify in a more specific manner which of the steps in the process match the standard criteria and which have discrepancies.

The application of the Discrepancy Evaluation Model in this study provides an alternative perspective to the existing literature on test items which has been primarily concerned with the quality of the test items as products. The model provides a means to evaluate the quality of test item development through its standards, readiness of the program, process, and product components. Therefore, this study not only tackles the issue of the quality of the resultant test items but also the issue of whether the process of test item

development adheres to the required quality assurance standards. This perspective is vital since the quality of the assessment products is determined by a number of interrelated processes.

Based on this research gap, this study aims to (1) evaluate the alignment of the implementation of Arabic School Exam item development at Muhammadiyah Elementary Schools in the Special Region of Yogyakarta with established standards based on the stages of the Discrepancy Evaluation Model (2) identify gaps between standards and actual conditions in the design, implementation, process, and product stages of test item development; (3) formulate recommendations for improvement to strengthen the quality assurance system for Arabic assessment development within the Muhammadiyah educational environment. With these objectives, this study is expected to broaden the scope of assessment evaluation from a focus predominantly on product quality toward an evaluation of the entire instrument development process, while also providing an empirical basis for strengthening the assessment quality assurance system in Muhammadiyah schools.

METHOD

This study is an evaluative research using a descriptive approach based on the Discrepancy Evaluation Model developed by Provus. This model was used to evaluate the alignment between standards and the implementation of test item development for the Arabic Language School Exam for sixth-grade students at Muhammadiyah elementary schools in the Special Region of Yogyakarta through five evaluation stages: design, installation, process, product, and comparison. The evaluation standards refer to the Guidelines for the Development of ISMUBA Test Items from the PWM DIY Council for Basic and Non-Formal Education, the ISMUBA Curriculum, and Ministry of Education and Culture Regulation No. 10 of 2025.

Research participants were selected using purposive sampling based on their direct involvement in the test item development program. The participants consisted of five individuals: one representative from the PWM DIY Council for Basic and Secondary Education and Non-Formal Education as the program administrator; two Arabic teachers from Muhammadiyah elementary schools who served as test item developers; and two principals of Muhammadiyah elementary schools whose teachers were selected as test item developers. The research instruments consisted of a questionnaire, an interview guide, and a document analysis sheet developed based on the indicators at each stage of the Discrepancy Evaluation Model. The content validity of the instruments was tested using the Content Validity Ratio (CVR) with the involvement of five validators, comprising experts in educational evaluation

and Arabic language subject matter experts. The reliability of the questionnaire was analyzed using Cronbach's Alpha coefficient with the aid of SPSS.

Questionnaire data obtained from the semantic differential scale were first transformed from an ordinal scale to an interval scale using the Method of Successive Intervals (MSI) before descriptive statistical analysis was conducted. Following this, the data was analyzed by calculating the achievement score and the percentage of the implementation for every indicator, categorizing it into five levels; very good (81%-100%), good (61%-80%), average (41%-60%), poor (21%-40%) and very poor (0%-20%). The data collected through interviews and document review was analyzed using three processes; data reduction, data presentation and concluding. All the analyses carried out were compared to the evaluation criteria to determine the level of compliance.

RESULTS AND DISCUSSION

Presentation of the research findings is made according to the Discrepancy Evaluation Model. This model comprises the following stages: design, installation, process, product, and synthesis of program discrepancies. The data collection has been carried out using a combination of questionnaires, interviews, and analysis of documents. It helps to find out the level of congruence between the criteria of creation of the Arabic School Exam questions and their implementation. The discussion not only explains the level of achievement at each stage but also interprets the significance of the findings, identifies the factors causing the discrepancies, relates the findings to previous research and perspectives on assessment quality assurance, and explains their implications for improving the test item development system.

The determination of evaluation categories in Table 1 is based on the results of a comparison between program standards and actual conditions obtained through data source triangulation to provide a comprehensive picture of the level of alignment in the implementation of Arabic School Exam question development.

Table 1. Evaluation Result Categories

Category	Determination Criteria
Compliant	Standards have been met based on the triangulation results
Needs Improvement	Some standards have been met, but weaknesses were still found in one or more data sources.
Not Met	The standards have not been met, or significant discrepancies have been found.

The categorization of these evaluation results is based on the fundamental principles of the Discrepancy Evaluation Model, which assesses the gap between standards and actual performance(Provus, 1971) , and the validation of findings is conducted through triangulation of data sources as developed in the qualitative research approach(Denzin, 2007) .

Design Phase

The evaluation results showed that all components in the design phase met the standards (see Table 2), so no discrepancies were found between actual conditions and program standards. Guidelines for question development, activity schedules, and supporting documents were available and were used as references. Question development also referred to the ISMUBA Curriculum and applicable guidelines, while the program objectives were understood as the basis for question development.

Table 2. Results of the Design Phase Evaluation

Aspect	Findings	Status
Planning Documents	Guidelines, schedules, and supporting documents	Compliant
ISMUBA Curriculum Development Standards	ISMUBA Curriculum, Development Guidelines	In accordance with
Program objectives	Understood as the basis for developing test items	In accordance

Interview results reinforce these findings by showing that all developers understood the question development guidelines as the primary reference during the instrument development process. A document review also showed that all required documents were available and used consistently during the planning stage. Thus, the triangulation results indicate that the design aspect met all established standards and was therefore categorized as “in accordance.”

From the perspective of the Discrepancy Evaluation Model, the design stage serves to establish standards that form the basis for comparison across all subsequent evaluation stages (Cutamora et al., 2025). Therefore, the quality of program implementation is significantly influenced by the clarity of objectives, program specifications, and the completeness of documents prepared during the planning stage (Aldabergenova et al., 2026). The findings of this study indicate that the gaps that emerge in subsequent stages are not caused by weaknesses in program design but rather by implementation factors.

This suggests that the quality assurance system during the planning stage has been implemented in accordance with quality assurance principles namely, ensuring that all procedures, standards, and program specifications are in place before operational activities begin. According to Downing and Haladyna (1997), the quality of assessment instruments is established as early as the planning stage through the development of test specifications, the definition of competencies to be measured, and the provision of clear guidelines for instrument development. Thus, the success of the design stage is a critical prerequisite for producing a systematic and well-documented instrument development process.

Furthermore, the findings from this study emphasize the notion that the quality of an assessment is dependent not only on the psychometric attributes of the instrument itself, but also on the quality of the system supporting the process of its development. As noted by Birenbaum (2007), “The validity of an assessment is substantiated using different sources of evidence including the consistency of the assessment’s goals, standards, and its development process”. The quality of an assessment requires transparency, standardization, and governance of the procedures starting from the planning stage (Divjak et al., 2023; Nash et al., 2026).

The existence of guidelines for item development, the curriculum and implementation timeline in this study signifies that these criteria have been met, thus ensuring a good starting point for the implementation of the program. Nonetheless, the successful completion of the design phase cannot be considered a guarantee of the program's successful implementation in its entirety. The future implementation of the program will depend on the organization’s ability to implement these standards into practice.

Installation Phase

The outcome of the evaluation conducted during the installation phase (Figure 1) reflected that the percentage achievement was 79.37% and thus categorized as “good.” The majority of indicators reached the category of “very good,” especially test blueprints availability, test development tools, the level of expertise of the test developers, supporting documents, and understanding of the guidelines. Nevertheless, some areas needing improvement were the number and competency of test developers, time allocation, supporting facilities, and standard operating procedures application.

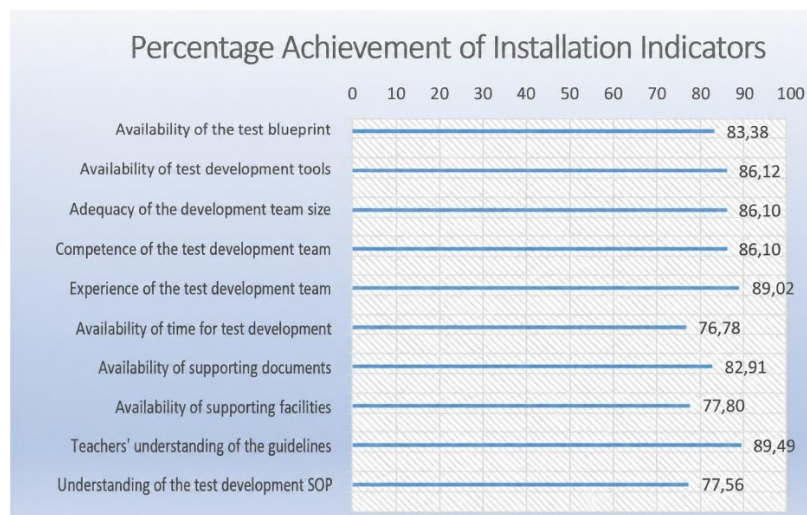


Figure 1. Installation Phase Survey Results

Verification of the results of the survey was made using interviews and document review to provide a better understanding of the compliance of implementation of the

installation phase with the existing standards. The triangulation results of the installation phase evaluation can be seen in Table 3.

Table 3. Triangulation Results of the Installation Phase Evaluation

Aspect	Findings	Status
Equipment Readiness	Test blueprints, guidelines, and supporting documents are available as references for question development	Compliant
Human Resource Readiness	These test constructors fulfill the criteria set out but their quantity, experience and competence require improvement	Needs Improvement
Implementation Support	There is inadequate time and facilities, as well as lack of implementation of the SOPs, that help in test item construction process	Needs Improvement
Understanding of guidelines	The guidelines for writing questions have been understood by the test constructors	Compliant

Based on the findings above, the primary challenge that comes up in the implementation stage is not the lack of administration tools, but the mismatch between the state of readiness in administration and operations. Although guidelines and blueprints for testing are present, the tools needed for implementation have not been completely developed yet. This situation indicates that a program's readiness for assessment quality assurance must be understood in a broader sense than merely the availability of documents.

Limitations in the number of test item developers and available time are critical findings because both factors are directly related to the opportunity to conduct quality control processes. Test item development is not merely the activity of writing test items, but involves a series of time- and labor-intensive activities, including understanding specifications, drafting test items, conducting reviews, discussing review results, making revisions, and documenting changes. If the number of test item developers is not proportional to the workload or if the available time is too limited, there is a risk that these stages will be carried out only minimally.

This finding aligns with Downing and Haladyna (1997), who characterize instrument development as a systematic process requiring adequate planning and resources. Instrument development requires stages that allow for review and refinement before the instrument is used (Ediyanto et al., 2022; Sarea et al., 2025). Thus, the time constraints identified in this study are not merely an administrative issue but can be a factor affecting quality assurance at both the process and product stages.

Findings regarding SOPs also have important implications. Knowledge of the guidelines is adequate, however, the implementation of SOPs is not yet optimal. This means that the knowledge about the standards is not yet transformed into the proper procedure follow up. With respect to quality assurance, this situation can be called an "implementation

gap” - the situation when standards exist and are known by practitioners, but they are not implemented yet. Thus, the assurance of high quality requires not only spreading of the guidelines, but also monitoring measures, which would control procedure follow up.

Thus, the “installation” stage becomes crucial for this research. The “gaps” on this stage may become the first stage of conditions explaining the emergence of the problems on the “process” stage. For example, limitations of the human resources and time can lead to the limitations in the coordination and communication, as well as to the increasing deadlines for revision and reviewing. This connection shows that the Discrepancy Evaluation Model allows not only to identify the gaps on each stage, but also to establish possible connections between them.

Process Stage

The result from the process stage evaluation emphasized the success in the implementation of item development, coordination among developers, the effectiveness of communication, item review and revision, the SOPs and schedules, as well as the completeness of documentation. From the survey result, the implementation of process stage had been accomplished with a level of completion rate of 64.63% which was in “good” category. While the highest achievement was related to the appropriateness of item development with the rate of 97.14%, the lowest one was the effectiveness of communication with the rate of 30.36%. The other indicators were in “ ” category including item review, item revision, and the compliance of SOPs (see Figure 2). However, coordination, schedule and the completeness of documentation still need attention.

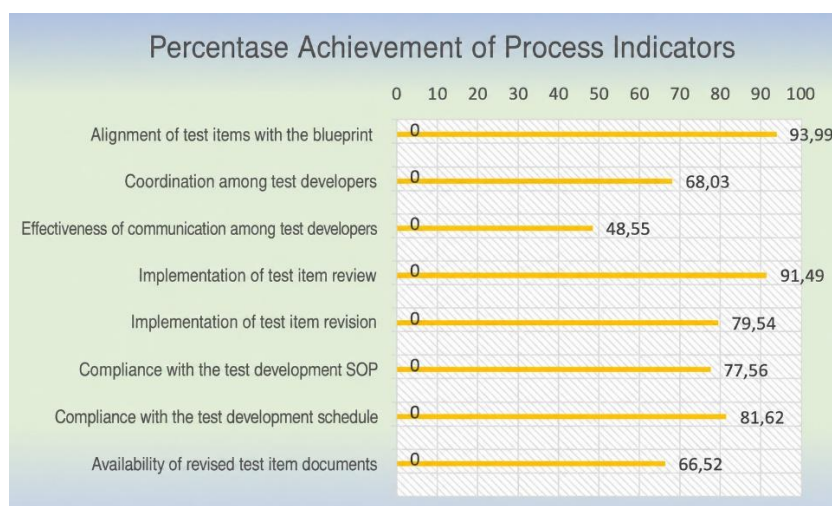


Figure 2. Process Stage Questionnaire Results

The results of the survey were validated by conducting interviews and document reviews in order to determine how well the process stage had been aligned with the existing standards. Triangulation results of the process stage evaluation can be seen in Table 4.

Table 4. Triangulation Results of the Process Stage Evaluation

Aspect	Findings	Status
Item Development	The writing, review, and revision of test items were carried out in accordance with the guidelines	In accordance
Coordination and communication	Coordination and communication among test item developers have not been optimal	Needs Improvement
Compliance with implementation	Implementation follows SOPs, but adherence to the schedule has not been consistent	Needs Improvement
Activity documentation	Administrative documents for test item development are incomplete	Needs Improvement

The main finding at this stage is the contrast between the technical quality of the test item writing and the quality of the governance of the development process. The 97.14% compliance rate in test item writing indicates that individual developers are relatively capable of translating the guidelines into test items. However, the communication effectiveness rate of only 30.36% indicates that these individual technical capabilities are not yet fully supported by an effective coordination system.

This discrepancy is significant because instrument development is, by its very nature, not solely an individual activity. Assessment quality requires a collaborative process of review, discussion, feedback, revision, and decision-making (Zhang, 2024). When communication among test developers is ineffective, there is a risk that issues that could be identified through peer review or discussion are not addressed optimally (Soares et al., 2021).

These findings expand upon previous research, which generally assessed item quality based on product characteristics. Lismawati's (2019) study, for example, revealed variations in difficulty levels and discriminative power in ISMUBA items. While these findings provide important information about product quality, this study demonstrates that issues regarding assessment quality can be traced further back to the processes that generate those products. In other words, weaknesses in the product are not always merely a matter of item-construction ability but may be related to how the coordination, review, and revision systems are managed.

Weaknesses in documentation were also a key finding. Documentation is not merely an administrative formality but is part of quality assurance because it provides evidence that a procedure has been carried out. A trace of audit trail consisting of review records, minutes, and revision documents will help managers track the decision-making process and changes made to the instrument. Lack of such documentation will make it harder for the organization

to assess the process and discover where improvements can be made in the next development cycle.

Such findings correspond with the Al Fraidan (2025) position on assessment quality, which states that procedural transparency and accountability are crucial aspects. "Therefore, the documentation deficiency observed in this case study can be perceived as a lack of the process accountability mechanism rather than an administration problem." According to "(Shafiee et al., 2022), procedural approach and documentation are vital when developing the content of an assessment.

The suboptimal adherence to the schedule can be also considered the result of installation problems. Time constraint identified in the previous phase can be carried forward and create pressure on the process phase. In conditions of time constraints, developers usually have priority in completing the product rather than making sure of the completion of coordination, documentation and review processes. Thus, the relationship between installation gap and process gap is revealed: resource readiness problems can create pressure for implementation, which affects the process governance in turn.

Therefore, although the process phase is generally rated in the "good" category, such aggregate numbers should be considered with a grain of salt. Good average score can hide weak indicators that are of particular importance, especially communication and documentation. From the perspective of the Discrepancy Evaluation Model, such weaknesses remain uncovered as they do not fall in the general category, but are compared with standards individually.

Product Stage

Based on the evaluation results of the product stage in Figure 3, there was an attainment rate of 77.66%, putting the evaluation into the "good" category. The quality of the product is rather high for manuscript and answer key archiving (100%), appropriateness of the content (97.96%) and language appropriateness (95.24%). However, there are still weaknesses regarding the timely accomplishment and question construction.

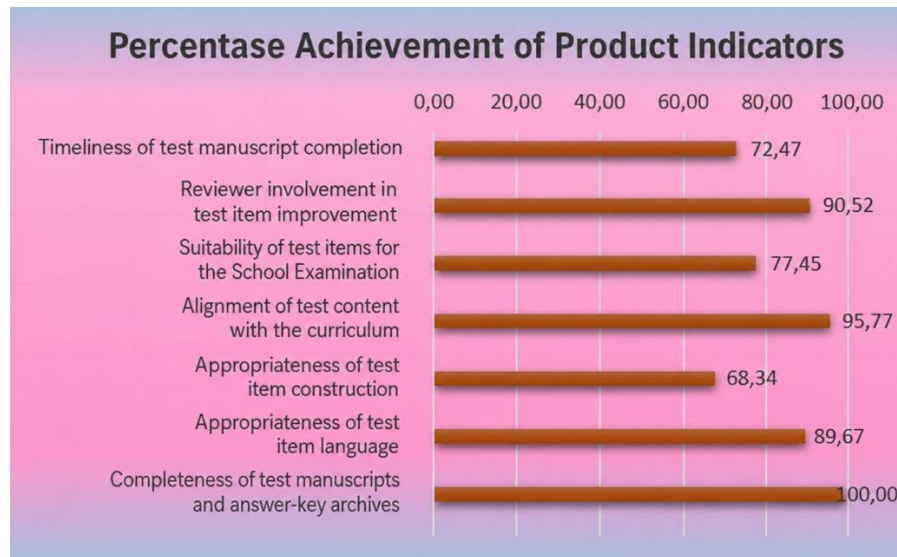


Figure 3. Product Phase Questionnaire Results

Subsequently, the survey findings were confirmed via interviews and documentation in order to establish the correlation between the materials developed and the set standards. Triangulation results of the product phase evaluation can be seen in Table 5.

Table 5. Triangulation Results of the Product Phase Evaluation

Aspect	Findings	Status
Quality of question content	Content, language, and appropriateness of the questions meet the standards	Compliant
Reviewer involvement	The question review process was conducted prior to finalization	Compliant
Product administration	Archives of test items and answer keys are well-maintained	In accordance with
Timeliness and question construction	Product delivery has not been entirely on time, and the design of some questions still needs improvement	Needs Improvement

Findings on product phase reveal that the final product is of quality but the development process for this product still experiences several difficulties. The high level of correlation between the content and language suggests that the developers of the test have been able to develop questions that are aligned with the competencies and content. The availability of complete archives of test items and answer keys also indicates that the management of the final product is relatively more organized compared to the documentation during the process stage. However, high achievement in the product does not mean that the entire quality assurance system is optimal. The fact that question construction still requires refinement indicates that some weaknesses in the process can carry over into the final product.

This shows that the product's success in terms of content and language has not been fully matched by success in terms of construction.

Delays in product completion also indicate that the quality of assessments has a time dimension. An instrument that is technically sound but completed past the deadline can cause problems during the exam implementation phase and reduce the time available for final checks. Therefore, timeliness should not be viewed merely as an administrative indicator, but as one of the conditions that support quality control (Er et al., 2019). The closer the product is completed to the time of use, the less time is available for final verification and corrections.

These findings reinforce the view that assessment quality is the result of an interaction between product quality and process quality. Birenbaum (2007) identifies the development process as one source of evidence supporting assessment validity, while Downing and Haladyna (1997) emphasize the importance of systematic development procedures. Therefore, the relatively good quality of the product in this study should not be interpreted as evidence that the entire system is already optimal, but rather as an indication that a system with a solid foundation can still be strengthened during the implementation phase.

Discrepancy Stage

A synthesis of the evaluation results reveals a relatively consistent pattern of gaps. The design stage met the standards, whereas gaps began to appear in the installation stage, became more apparent in the process stage, and some of their effects were visible in the product stage. This shows that the problem is not one of design but one of the ability of the program to transform standards into practice. The synthesis matrix of discrepancies in test item development can be seen in Table 6.

Table 6. Synthesis Matrix of Discrepancies in Test Item Development

Stage	Location of Gap	Dominant Factor	Impact	Recommendations
Design	None	Thorough planning	Standards serve as a guide for implementation	Maintained
Installation	Human resources and Time	Limited personnel and time allocation	Program readiness is not yet optimal	Adding staff, rescheduling
Process	Coordination, communication, and documentation	Implementation management	Decreased effectiveness in question development	SOPs for coordination and documentation
Product	Question construction and language review	Impact of process weaknesses	Technical quality of test items is not yet optimal	Strengthening quality control

As seen from the matrix, discrepancies between stages are interrelated and dependent. Time and human resource constraints that may arise at the installation stage may influence the intensity of coordination and review at the process stage. Also, inappropriate coordination and review procedures might contribute to the sustainability of the weakness of item construction at the product stage. This is an important finding: the product discrepancies in this study cannot be distinguished from process discrepancies and program readiness.

It illustrates the principle of the Discrepancy Evaluation Model which states that evaluation should go further than determining compliance of the product with the criteria and should go back to the stage where a discrepancy arose (Kumar & Nti, 1998). Evaluation of the final product allows detecting the weakness of the test items construction but not understanding the reasons of these weaknesses. Through the discrepancy approach used in this research it is shown how the issues of product can be traced back to the readiness conditions and development process.

Also, these findings contribute to the literature on assessment quality assurance. In previous studies devoted to the topic of Arabic language assessment in Muhammadiyah context, attention was paid to item quality or implementation of the assessment procedure. The present study contributes to this research area through the demonstration that quality assurance has to be viewed not only as the final-stage review but as a tool working throughout the whole process of instrument development.

Substantively, research findings show that there is an excellent normative background of the process of developing Arabic School Exam items at Muhammadiyah Elementary Schools in DIY but the implementation capacity is not yet developed enough. That is why the “design” stage is evaluated as “compliant”, and “installation” and “process” stages need improvements. It means that the problem lies not in creation of new criteria but in their implementation under the operational conditions.

These research findings lead to the implication of the improvement of improvement strategies which are directed at the implementation issue. Firstly, the size and composition of the test development team have to be adapted to the amount of work so that writing, reviewing, and revising are not done by a few people only. Secondly, the time needs to be allocated realistically as the time devoted to the coordination, review, revision, and checking of the tests is needed to be considered. Thirdly, SOPs have to be transformed to work mechanism which include the responsible person, schedule, implementation proof, and monitoring mechanism. Lastly, documentation like the meeting minutes, official documents, and revision history have to be seen as quality assurance rather than administration formality.

In conclusion, the research findings show that the enhancement of the quality of assessment in the Arabic language for the Muhammadiyah context cannot be done only through the increase of the teacher's capability in writing the items of the test. The improvement also has to take account of the system, especially the resources, coordination, communication, documentation, schedule, and quality control. Therefore, the use of the Discrepancy Evaluation Model in this study does not only function as gap identification but also to identify the most strategic intervention points.

CONCLUSION

The findings from this study reveal that the development of questions for the Arabic School Exam in Muhammadiyah Elementary Schools in the Special Region of Yogyakarta has been carried out well according to the set standards, although some gaps have been found in program implementation. Using the Discrepancy Evaluation Model, it is found that the design phase meets all the standards, but the installation phase, process phase, and the product phase still need improvement in the aspects of resource readiness, management of implementation, coordination, documentation, and quality control of test items.

These findings show that the quality of the products of the assessment process is determined not only by the existence of the guidelines and the appropriateness of the materials but also by the effectiveness of the management system during the process of development of the instruments. Thus, it needs to be improved through increasing the ability of the test item development team, good time management and better mechanisms of coordination and documentation. Through the improvement, the development process of the questions for the Arabic School Exam will produce assessment instruments that are of high quality and more valid.

REFERENCES

- Aithal, A., & Aithal, P. S. (2020). Development and validation of survey questionnaire & experimental data: A systematical review-based statistical approach. *International Journal of Management, Technology, and Social Sciences*, 5(2), 233–251. <https://doi.org/10.47992/IJMTS.2581.6012.0116>
- Al Fraidan, A. (2025). Procedural Transparency and Legal Accountability to Sustain AI-Mediated Language Assessment in Saudi Arabia. *Sage Open*, 15(4), 21582440251396113. <https://doi.org/10.1177/21582440251396113>
- Aldabergenova, S., Alshynbayeva, Z., Jumadilova, N., Bogacz, P., & Akizhanova, A. (2026). Research governance and quality assurance measurement in Kazakhstan's higher education institutions: An indicator-based framework aligned with ISO 21001:2025 and ESG 2015 (EHEA). *Iberoamerican Journal of Science Measurement and Communication*, 6, 1–11. <https://doi.org/10.47909/ijsmc.355>

- Appels, L., De Maeyer, S., Faddar, J., & Van Petegem, P. (2022). Capturing quality. Educational quality in secondary analyses of international large-scale assessments: A systematic review. *School Effectiveness and School Improvement*, 33(4), 629–668. <https://doi.org/10.1080/09243453.2022.2115519>
- Babinčáková, M., Ganajová, M., Sotáková, I., & Bernard, P. (2020). Influence of Formative Assessment Classroom Techniques (Facts) on Student's Outcomes in Chemistry at Secondary School. *Journal of Baltic Science Education*, 19(1), 36–49. <https://doi.org/10.33225/jbse/20.19.36>
- Birenbaum, M. (2007). Evaluating the assessment: Sources of evidence for quality assurance. *Studies in Educational Evaluation*, 33(1), 29–49. <https://doi.org/10.1016/j.stueduc.2007.01.004>
- Chambers, R. L., Lahuerta-Martín, S., Greco, A. M., Pattinson, R., Pickles, T., Hassan, J., & Mookink, L. B. (2026). Assessing content validity: Challenges of conducting systematic reviews of patient-reported outcome measures and recommendations to improve the application of COSMIN guidance. *Quality of Life Research*, 35(6), Article 157. <https://doi.org/10.1007/s11136-026-04261-5>
- Cutamora, U., Leonardo, G., & Lagria, M. M. (2025). Evaluating a community-based rehabilitation program using Provus discrepancy evaluation model. *Recoletos Multidisciplinary Research Journal, Special Issue*, 1–19.
- Denzin, N. K. (2007). Triangulation. In G. Ritzer (Ed.), *The Blackwell encyclopedia of sociology*. Wiley. <https://doi.org/10.1002/9781405165518.wbeost050>
- Divjak, B., Svetec, B., Horvat, D., & Kadoić, N. (2023). Assessment validity and learning analytics as prerequisites for ensuring student-centred learning design. *British Journal of Educational Technology*, 54(1), 313–334. <https://doi.org/10.1111/bjet.13290>
- Downing, S. M., & Haladyna, T. M. (1997). Test Item Development: Validity Evidence From Quality Assurance Procedures. *Applied Measurement in Education*, 10(1), 61–82. https://doi.org/10.1207/s15324818ame1001_4
- Ediyanto, E., Sunandar, A., Ramadhani, R. S., & Aqilah, T. S. (2022). Sustainable Instrument Development in Educational Research. *Discourse and Communication for Sustainable Education*, 13(1), 37–47. <https://doi.org/10.2478/dcse-2022-0004>
- Er, H. M., Nadarajah, V. D., Hays, R. B., Yusoff, N., & Loh, K. L. Y. (2019). Lessons Learnt from the Development and Implementation of an Online Assessment System for Medical Science Programmes. *Medical Science Educator*, 29(4), 1103–1108. <https://doi.org/10.1007/s40670-019-00802-5>
- Fikriyah, N. (2021). Analisis Butir Soal Ulangan Tengah Semester Mata Pelajaran Bahasa Arab Kelas VII Semester Ganjil SMP Muhammadiyah 1 Yogyakarta Tahun Ajaran 2019/2020. *Maharaat: Jurnal Pendidikan Bahasa Arab*, 3(2), 128–140. <https://doi.org/10.18196/mht.v3i2.10501>
- Harris, D. J., Welch, C. J., & Dunbar, S. B. (2024). In the beginning, there was an item.... *Educational Measurement: Issues and Practice*, 43(4), 40–45. <https://doi.org/10.1111/emip.12647>

- Ing, M., Kosko, K. W., Jong, C., & Shih, J. C. (2024). Validity evidence of the use of quantitative measures of students in elementary mathematics education. *School Science and Mathematics*, 124(6), 411–423. <https://doi.org/10.1111/ssm.12660>
- Kumar, R., & Nti, K. O. (1998). Differential Learning and Interaction in Alliance Dynamics: A Process and Outcome Discrepancy Model. *Organization Science*, 9(3), 356–367. <https://doi.org/10.1287/orsc.9.3.356>
- Lismawati, L. (2019). Analisis Butir Soal Ismuba (Al-Islam Kemuhammadiyah dan Bahasa Arab) Kelas IV SD Muhammadiyah 5 Jakarta Selatan. *Jurnal Pendidikan Islam*, 10(1), 43–51. <https://doi.org/10.22236/jpi.v10i1.3452>
- Lucander, H., & Christersson, C. (2020). Engagement for Quality Development in Higher Education: A Process for Quality Assurance of Assessment. *Quality in Higher Education*, 26(2), 135–155. <https://doi.org/10.1080/13538322.2020.1761008>
- Mustafa, P. S. (2021). Model discrepancy sebagai evaluasi program pendidikan. *PALAPA*, 9(1), 182–198. <https://doi.org/10.36088/palapa.v9i1.1067>
- Nash, B., Karvonen, M., Flowers, C., Clark, A., Romine, R. S., & Bechard, S. (2026). A Comprehensive Alignment Framework for Building Coherence in an Assessment System. *Educational Measurement: Issues and Practice*, 45(2), e70027. <https://doi.org/10.1111/emip.70027>
- Otaya, L. G., Kartowagiran, B., & Retnawati, H. (2020). The construct validity and reliability of the lesson plan assessment instrument in primary schools. *Jurnal Prima Edukasia*, 8(2), 126–134. <https://doi.org/10.21831/jpe.v8i2.33275>
- Patil, R. P., Bahekar, S. E., Kulkarni, M. D., & Baig, M. S. (2022). Evaluation of validity and reliability of multiple-choice questions in second MBBS competency-based medical education-based pharmacology examination of medical institute of India. *International Journal of Research in Medical Sciences*, 10(12), 2878–2883. <https://doi.org/10.18203/2320-6012.ijrms20223091>
- Pepper, D. (2020). When Assessment Validation Neglects Any Strand of Validity Evidence: An Instructive Example from PISA. *Educational Measurement: Issues and Practice*, 39(4), 8–20. <https://doi.org/10.1111/emip.12380>
- Phelps, G., Gitomer, D. H., Iaconangelo, C. J., Etkina, E., Seeley, L., & Vokos, S. (2020). Developing Assessments of Content Knowledge for Teaching Using Evidence-centered Design. *Educational Assessment*, 25(2), 91–111. <https://doi.org/10.1080/10627197.2020.1756256>
- Provus, M. M. (1971). *Discrepancy evaluation: For educational program improvement and assessment*. McGraw-Hill.
- Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. (2024). Item analysis: The impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, 24(1), 445. <https://doi.org/10.1186/s12909-024-05433-y>

- Robbany, M. J., Erlina, E., & Hijriyah, U. (2025). Literature Analysis of Arabic Language Evaluation Instrument Strategies and Designs in the Context of CBT. *EL-IBTIKAR: Jurnal Pendidikan Bahasa Arab*, 14(2), 305–311. <https://doi.org/10.24235/ibtikar.v14i2.20143>
- Sarea, M. S., Dewantara, A. H., & Sabbar, S. D. (2025). Development of A Comprehensive Instrument for The Islamic Education Subject Examination. *Tafkir: Interdisciplinary Journal of Islamic Education*, 6(3), 756–767. <https://doi.org/10.31538/tijie.v6i3.1940>
- Shafiee, M., Shanbehzadeh, M., Nassari, Z., & Kazemi-Arpanahi, H. (2022). Development and evaluation of an electronic nursing documentation system. *BMC Nursing*, 21(1), 15. <https://doi.org/10.1186/s12912-021-00790-1>
- Soares, S. F., Carvalho Moura, E. C., Lopez, V., & Peres, A. M. (2021). Professional Nursing Communication Competence: Theoretical procedures for instrument development and pilot test. *Journal of Nursing Management*, 29(6), 1496–1507. <https://doi.org/10.1111/jonm.13283>
- Sosa Paredes, Y. K., & Andersson, B. (2025). Validity of test score interpretations and cross-cultural comparisons in the First and Second International Science Studies. *Educational Assessment, Evaluation and Accountability*, 37(2), 231–254. <https://doi.org/10.1007/s11092-024-09444-7>
- Qalbi, T., Baso, Y. S., & T., Y. (2026). Development of Arabic language learning outcome assessment instrument based on Hot Potatoes application at junior high school. *Jurnal Pendidikan Dasar dan Keguruan*, 11(1), 74–81. <https://doi.org/10.47435/jpdk.v11i01.4376>
- Twist, L., & Fraillon, J. (2020). Assessment Content Development. In H. Wagemaker (Ed.), *Reliability and Validity of International Large-Scale Assessment* (Vol. 10, pp. 37–59). Springer International Publishing. https://doi.org/10.1007/978-3-030-53081-5_4
- Velasco Fajardo, R. (2023). Assessment Tools Used by the Senior High School Mathematics Teachers during Modular Distance Learning. *International Journal of Science and Research (IJSR)*, 12(6), 470–474. <https://doi.org/10.21275/SR23603193125>
- Wang, L., Chen, J., & Talib, O. (2025). Core dimensions and measurement characteristics of multidimensional nursing competency assessment instruments: A systematic review. *Journal of Professional Nursing*, 61, 96–104. <https://doi.org/10.1016/j.profnurs.2025.09.013>
- Wasito, W. (2019). Implementasi Kurikulum Ismuba di SD Muhammadiyah Karangkajen Yogyakarta. *Jurnal Pendidikan Islam*, 10(1), 1–18. <https://doi.org/10.22236/jpi.v10i1.3447>
- Zhang, Q. (2024). Design and research of serial communication system based on virtual instrument technology. *Journal of Physics: Conference Series*, 2853(1), 012023. <https://doi.org/10.1088/1742-6596/2853/1/012023>