Lughawiyah Journal of Arabic Education and Linguistics, 7(2) - December 2025 88-112

Lughawiyah
Journal of Arabic Education and Linguistics
E-ISSN 2715-8101
P-ISSN 2715-8098

Development and Validation of a CBT-Based Arabic Question Bank for Secondary Education

Nurul Aini Pakaya[1], Ibnu Rawandhy N. Hula[2], Nikma S. Abdjul[3], Adtman A. Hasan[4] Yousif Abdelmannan Mohamed Godat Arrashedy[5]

[1] *Universitas Muhammadiyah Gorontalo*
[2] *Institut Agama Islam Negeri Sultan Amai Gorontalo, Indonesia*
[3] *Institut Agama Islam Negeri Sultan Amai Gorontalo, Indonesia*
[4] *Institut Agama Islam Negeri Sultan Amai Gorontalo, Indonesia*
[5] *International Institute of Khartoum, Sudan*

Corresponding Author: Nurul Aini Pakaya, Email: nurulainipakaya@umgo.ac.id

Received: August 19, 2025	Revised: Sep 22, 2025	Accepted: Nov 25, 2025	Online: Dec 31, 2025

ABSTRACT

The educational process often overlooks the evaluation aspect of learning, as educators frequently develop assessment instruments without adhering to established guidelines, despite evaluation being a critical component in determining the effectiveness and development of education. This study employed a mixed-methods Research and Development (R&D) approach to develop an Arabic question bank for Grade X students at State Islamic High School (MAN) 1 Pohuwato. Quantitatively, logical validity results showed that expert validators awarded a total score of 74 with an average of 3.89, while practitioner validators provided a total score of 71 with an average of 3.74, both categorized as very valid, resulting in a combined average validity score of 3.815. Qualitatively, the high validity values were supported by expert and practitioner evaluations indicating that the test items were clearly formulated, aligned with curriculum competencies, and appropriately represented the intended cognitive levels. Following the validation process, the revised items were tested in two classes, X-A and X-D, involving 36 students. This development research produced a set of 40 Arabic multiple-choice questions ready for use in the Mid-Term Examination (UTS) and Final Examination (UAS). The implications of this study show that CBT-based Arabic question banks can improve the objectivity, efficiency, and quality of Arabic learning evaluation in madrasas.

Keywords: *Arabic, Computer Bassed Test, Question Bank Development*

Journal Homepage https://ejournal.uinmybatusangkar.ac.id/ojs/index.php/lughawiyah

How to cite: Pakaya, A. N., Hula, N. R. I., Abdjul, S. N., Hasan, A. A., & Arrashedy, G. M. A. Y. (2025). Development and Validation of a CBT-Based Arabic Question Bank for Secondary Education. *Lughawiyah Journal of Arabic Education and Linguistics*, 7(2), 88-112. http://dx.doi.org/10.31958/lughawiyah.v7i2.16058

Published by: Universitas Islam Negeri Mahmud Yunus Batusangkar, Indonesia

INTRODUCTION

Assessment and evaluation constitute essential components of the educational process; however, in practice, they often receive insufficient attention from teachers and lecturers, particularly in language learning contexts (Hastini et al., 2020). Many educators still design evaluation instruments pragmatically without strictly following established procedures for test construction, such as alignment with learning objectives, validity, reliability, and item analysis (Moh Firdaus et al., 2023). This condition poses a serious challenge because improperly developed assessment tools may lead to inaccurate measurement of students' competencies and flawed educational decision-making regarding the effectiveness of instruction (Rayón et al., 2014).

Education is inseparable from assessment practices, as assessment serves not only to measure learning outcomes but also to evaluate the success of instructional processes and inform continuous improvement (Mohamad Aso Samsudin & Ukhtul Iffah, 2019). High-quality assessment systems are widely recognized as a key factor in improving educational quality because they allow educators to identify learners' strengths and weaknesses and to refine pedagogical strategies accordingly (Natadireja & Nurachadijat, 2023). In foreign language education, particularly Arabic language learning, assessment plays a crucial role due to the complexity of linguistic competencies involved, including vocabulary, grammar, reading comprehension, and contextual understanding. Therefore, evaluation instruments must be developed carefully to ensure that they accurately reflect students' actual language proficiency (Gebril, 2017).

A test, as a formal assessment instrument, is designed to measure students' understanding using a set of structured questions or tasks (Elfira et al., 2023). The quality of a test is primarily determined by its validity and reliability, which indicate the extent to which the instrument measures what it is intended to measure and produces consistent results (Desiriah & Setyarsih, 2021). In Arabic language learning, good test items should align with curriculum objectives, measure the intended competencies, and correspond to students' cognitive levels (Noval & Adhani, 2021). However, empirical observations at State Islamic High School (MAN) 1 Pohuwato reveal that Arabic evaluation practices often rely on hastily prepared test items, including the reuse of questions obtained from online sources without systematic validation. Such practices not only reduce the credibility of assessment results but also increase the likelihood of academic dishonesty and inaccurate depiction of students' actual abilities.

In terms of implementation, conventional paper-based testing remains dominant at State Islamic High School (MAN) 1 Pohuwato. This assessment method presents several limitations, including high duplication costs, time-consuming scoring procedures, and limited test security. These conditions contradict one of the fundamental characteristics of a good test, namely efficiency and cost-effectiveness. Moreover, cheating during examinations is still frequently encountered, undermining the objectivity and fairness of assessment outcomes. In the context of rapid developments in science and technology in the 21st century, such limitations highlight the urgent need for educational institutions to adopt more modern and technology-based assessment systems. Arabic language teachers at State Islamic High School 1 Pohuwato, together with the Arabic MGMP team in Pohuwato Regency, have long expressed a desire to implement computer-based examinations to improve the quality, efficiency, and integrity of learning evaluations.

Computer-Based Test (CBT) systems have been increasingly adopted in educational assessment as an effective alternative to conventional testing methods. According to Syarifah and Asda (2021) CBT is generally used to enhance the efficiency and effectiveness of test administration, while also improving test security and scoring objectivity. CBT refers to examinations conducted using computers, eliminating the need for paper, pen, or pencil, as questions and answer sheets are presented digitally (Luecht, 2015). The utilization of CBT has been shown to reduce operational costs, save time, and facilitate faster feedback for both teachers and students (Mujiatun et al., 2022). International studies further indicate that computer-based assessment can significantly enhance students' engagement and motivation, particularly when interactive digital interfaces are employed (Pellas, 2025).

Alongside the development of CBT systems, the establishment of structured question banks has become an important innovation in educational assessment. A question bank is a systematically organized collection of validated test items that can be reused while maintaining consistency, fairness, and alignment with curriculum standards (Beerepoot, 2023). Research Krzic & Brown, (2022) demonstrates that well-developed question banks contribute to more effective and reliable assessment practices.

Nevertheless, studies that specifically focus on the development of Arabic language question banks integrated with Computer-Based Test (CBT) systems remain limited, particularly at the secondary education level. Existing research on CBT-based assessment has largely concentrated on implementation effectiveness, technological readiness, or comparative performance outcomes, predominantly within English language assessment contexts, while the systematic development and empirical validation of Arabic test items are often treated as separate or secondary concerns. This indicates a clear research gap, as few studies have simultaneously addressed curriculum-aligned Arabic question bank construction, expert and practitioner validation, and comprehensive psychometric analysis, including item validity, reliability, difficulty index, and discrimination power, within a CBT-based evaluation framework (Malik & Malik, 2020).

This study positions itself within this research gap by developing an Arabic question bank integrated with a CBT application for Grade X students at State Islamic High School (MAN) 1 Pohuwato. The novelty of this research lies in its comprehensive approach, combining systematic test item development with empirical psychometric analysis, including validity, reliability, difficulty level, and discrimination power testing. Unlike previous studies that emphasize technological implementation alone, this study ensures that the digital assessment system is supported by high-quality, empirically validated test items. This integrated approach aligns with contemporary assessment frameworks that emphasize both technological innovation and measurement accuracy (Punoševac & Nikolić, 2024).

The significance of this research is both practical and theoretical. Practically, the developed Arabic question bank provides teachers with a validated and reliable assessment instrument that can be implemented efficiently through a CBT system, thereby reducing costs, minimizing cheating, and improving scoring objectivity. Theoretically, this study contributes to the existing body of knowledge on Arabic language assessment by providing empirical evidence on the effectiveness of integrating question bank development with computer-based testing in a madrasah context. Such contributions are particularly relevant for developing countries, where digital transformation in education is still uneven and underexplored in empirical research (Bahrun et al., 2023; Makruf & Barokah, 2023)

Previous studies on CBT implementation have predominantly focused on English language assessment, particularly in examining students' performance, test anxiety, and score comparability between paper-based and computer-based formats (Yu & Iwashita, 2021). These studies generally emphasize the effectiveness and practicality of CBT platforms rather than the psychometric quality of the test instruments themselves (Reza Amirian et al., 2023). In contrast, research on CBT-based assessment in Arabic language education remains limited, especially studies that integrate systematic question bank development, expert validation, and higher-order thinking skills (HOTS) orientation (Hakami et al., 2015; Strother et al., 2008) Therefore, this study addresses a critical research gap by developing and validating an Arabic question bank designed specifically for CBT-based UTS and UAS, thereby extending existing CBT assessment research beyond English language contexts and contributing original insights to Arabic language evaluation practices.

Based on these considerations, this study is guided by the hypothesis that an Arabic question bank developed through a systematic Research and Development (R&D) approach and implemented using a CBT system will demonstrate high levels of validity and reliability and will be feasible for use as an evaluation instrument in Arabic language learning. The primary variables investigated in this study include the quality of test items (validity, reliability, difficulty index, and discrimination power) and the feasibility of CBT-based implementation. Methodologically, this research employs a Research and Development (R&D) design, utilizing expert validation, practitioner validation, and field trials to ensure the empirical quality of the developed product.

The novelty of this study lies in its integrative approach, which not only implements CBT for Arabic language assessment but also systematically develops and empirically validates a curriculum-based Arabic question bank with comprehensive psychometric analysis, including item validity, reliability, difficulty index, and discrimination power—an aspect that has been largely overlooked in previous CBT studies, particularly those focused on English language assessment. Through these objectives, this study seeks to provide an innovative, empirically grounded contribution to the advancement of Arabic language assessment and digital evaluation practices in secondary education.

Therefore, the main objective of this study is to develop and validate an Arabic question bank using a CBT application for Grade X students at State Islamic High School (MAN) 1 Pohuwato. Specifically, this research aims to (1) design Arabic test items aligned with curriculum competencies, (2) evaluate the validity and reliability of the developed items, and (3) assess the feasibility of implementing the question bank through a CBT system. Through these objectives, this study seeks to provide an innovative, empirically grounded contribution to the advancement of Arabic language assessment and digital evaluation practices in secondary education.

RESEARCH METHODOLOGY

Research Design and Approach

This study employed a Research and Development (R&D) approach aimed at developing, validating, and testing an Arabic question bank integrated with a CBT application (Nugroho et al., 2022). The R&D method was selected because it is appropriate for producing educational products and evaluating their feasibility and effectiveness through systematic stages (Muttaqin, 2019). The research was conducted at at State Islamic High School (MAN) 1 Pohuwato during the academic year 2022/2023.

Research Subject and Location

The population of this study consisted of all Grade X students of at State Islamic High School (MAN) 1 Pohuwato. The research sample was selected using a purposive sampling technique, involving two classes, namely class X-A and class X-D, with a total of 36 students. These classes were chosen based on their representativeness and readiness to participate in CBT-based assessment trials. Arabic teachers at at State Islamic High School (MAN) 1 Pohuwato and one Arabic language assessment expert were also involved as validators in this study.

Data Collection Instruments and Techniques

The development model adopted in this research was adapted from the 4-D model proposed by Thiagarajan, Semmel, and Semmel, which includes Define, Design, Develop, and Disseminate stages. However, due to the practical limitations of the research scope, this study implemented only three stages, namely Define, Design, and Develop, while the Disseminate stage was not conducted (Utomo, 2021). The Define stage involved needs analysis and identification of assessment problems in Arabic learning. The Design stage focused on constructing test blueprints, item grids, scoring guidelines, and CBT system integration. The Develop stage consisted of expert validation, practitioner validation, product revision, and field trials.

The research instruments included (1) an Arabic question bank consisting of 40 multiple-choice items, (2) validation questionnaires for expert and practitioner assessments, and (3) a CBT application used to administer the test. The validation questionnaire was designed using a Likert scale ranging from 0 to 4, representing very invalid to very valid criteria. The assessment aspects covered content relevance, construct accuracy, linguistic clarity, and alignment with curriculum competencies (Hatamnejad et al., 2023).

Table 1. Details of Research Instruments

No	Research Instruments	Detailed Description	Measured Aspects/Dimensions	Scale & Interpretation
1	Arabic Question Bank (40 PG items)	The developed question bank contains 40 multiple-choice items designed based on the Arabic class X learning grid. The question items reflect the three core themes of the curriculum, measuring knowledge, understanding, and high-level thinking skills according to the revised Bloom taxonomy.	- Compatibility with KD & curriculum indicators - Representation of Arabic material - Linguistically valid question structure	40 item PG terprogram untuk evaluasi UTS / UAS berbasis CBT & tertulis
2	Expert & Practitioner Validation Questionnaire	The questionnaire is used to measure the quality of the question bank through the assessment of experts and practitioners (Arabic teachers). The instrument is designed based on the Likert scale of 0–4 with scale points: 0 = very invalid, 1 = invalid, 2 = quite valid, 3 = valid, 4 = very valid.	- Content relevance - Construct accuracy - Linguistic clarity - Curriculum fit	Scale 0–4: Total score maps the validity of the instrument
3	Aplikasi CBT (Computer-	An application that is used as a medium to digitally administer questions to students.	- Integration of technology in	Automatic score per item

Based Test)	This application allows timing, automatic scoring, and generating empirical data for analysis of reliability, difficulty, and power of discriminatory questions.	evaluation - Ease of exam administration - Student response log data	→ statistically tested data

The research procedures were conducted sequentially. Initially, the developed test items and instruments were reviewed by an Arabic language assessment expert and Arabic teachers as practitioners. Based on the validators' feedback, revisions were made to improve item clarity and instructional accuracy. Subsequently, the revised question bank was tested through two field trials involving the same student sample using the CBT system. Each trial was conducted within a 40-minute test duration to simulate actual midterm and final examination conditions.

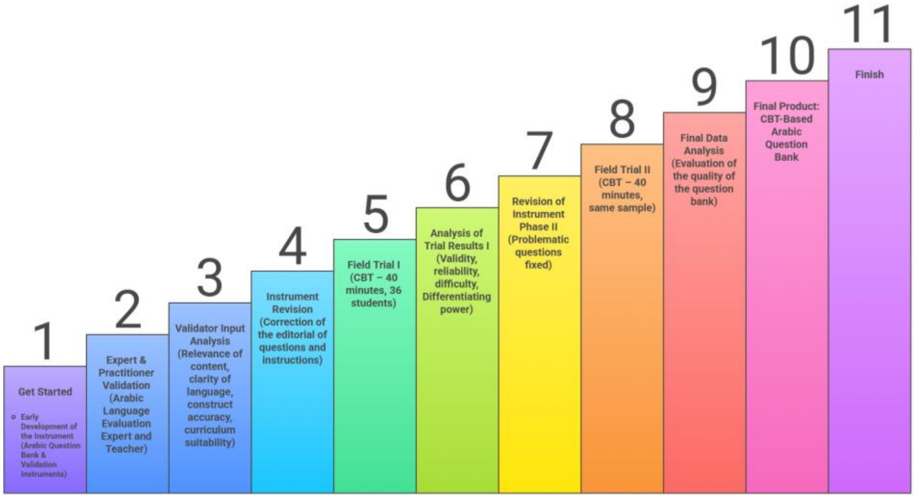
1
Get Started
Early Development of the Instrument (Arabic Question Bank & Validation Instruments)
2
Expert & Practitioner Validation (Arabic Language Evaluation Expert and Teacher)
3
Validator Input Analysis (Relevance of content, clarity of language, construct accuracy, curriculum suitability)
4
Instrument Revision (Correction of the editorial of questions and instructions)
5
Field Trial I (CBT – 40 minutes, 36 students)
6
Analysis of Trial Results I (Validity, reliability, difficulty, Differentiating power)
7
Revision of Instrument Phase II (Problematic questions fixed)
8
Field Trial II (CBT – 40 minutes, same sample)
9
Final Data Analysis (Evaluation of the quality of the question bank)
10
Final Product: CBT-Based Arabic Question Bank
11
Finish

Figure 1. Flowcharts of the Research Process

Data Analysis Techniques

Data analysis in this study employed both qualitative and quantitative techniques. Qualitative data were obtained from validators' comments and suggestions, which were used to revise the test items. Quantitative data were analyzed using descriptive and inferential statistics with the assistance of SPSS version 26. Item validity was examined using biserial correlation, where items were considered valid if the calculated correlation coefficient exceeded the critical value at a 0.05 significance level. Test reliability was measured using Cronbach's Alpha coefficient. Item difficulty levels and discrimination indices were also calculated to determine the quality and effectiveness of each question item.

To ensure the validity and reliability of the research findings, several strategies were implemented. Content validity was established through expert and practitioner validation. Empirical validity and reliability were confirmed through field testing and statistical analysis. The use of multiple indicators, repeated trials, and standardized scoring procedures further strengthened the credibility of the developed assessment instrument.

The methodology of this study was limited to the development and testing of an Arabic question bank for Grade X students at a single madrasah, namely State Islamic High School (MAN) 1 Pohuwato. Therefore, the findings may not be generalized to other grade levels or educational contexts without further testing. Additionally, the study focused on multiple-choice test items and did not include other assessment formats such as essays or performance-based tasks. Despite these limitations, the methodology provides a systematic and replicable framework for developing CBT-based assessment instruments in Arabic language learning.

RESULT

This section presents the empirical findings of the development and validation of the Arabic question bank and discusses their meaning, significance, and implications for CBT-based Arabic language assessment. The discussion explicitly connects the findings to the research objectives and the broader discourse on assessment quality, item validity, and higher-order thinking skills (HOTS) evaluation.

Development Results of the Arabic Question Bank

In the planning stage, researchers have prepared preliminary designs or sketches of question bank products used in CBT-based Mid-Term Examination/Final Examination at State Islamic High School (MAN) 1 Pohuwato. This stage is crucial because it determines the alignment between learning objectives, assessment formats, and the technological platform used for delivery. This process involves determining the instrument's shape, preparing the grid, and designing the instrument.

The planning stage incorporated explicit mapping from curriculum standards to item blueprints, ensuring that each item could be traced to a specific learning indicator and cognitive level; this mapping served as the primary quality-control mechanism before item writing. Such systematic blueprinting reduces the risk of construct underrepresentation and facilitates later psychometric analyses.

From a development perspective, this stage reflects the principle that assessment instruments must be intentionally designed rather than merely compiled, particularly when the goal is to measure higher-order cognitive abilities.

Instrument Shape Assignment

In this study, the evaluation tool chosen to measure the higher-order thinking skills of grade X learners in Arabic subjects was a written test in the form of multiple choice. The selection of a multiple-choice format was a strategic methodological decision rather than a matter of convenience.

First, multiple-choice tests can measure various cognitive levels, from the level of memory to the ability to create. When well-constructed, multiple-choice items can assess analysis, evaluation, and reasoning, which are core indicators of HOTS. Therefore, this test is appropriate for evaluating learners' higher-order thinking skills.

However, it should be noted that crafting multiple-choice items for HOTS requires careful item-writer training and use of plausible distractors that reflect common misconceptions; without these features, multiple-choice items tend to assess lower-order recall. Consequently, the development process included item-writer guidance on constructing stems that require inference, interpretation, and evaluation rather than rote recall (Alkhatib, 2022).

Second, the assessment of this test is carried out efficiently, quickly, and objectively, making it suitable for researchers with time constraints. This efficiency is particularly important in CBT environments where automated scoring is required.

Third, multiple-choice tests can cover a wide range of material because researchers include four essential competencies in preparing questions. This breadth allows the instrument to represent the curriculum proportionally and avoid construct underrepresentation.

Creating a problem item framework is a crucial aspect of the design process. The problem framework to be prepared refers to the 2013 curriculum or the applicable curriculum. The purpose of creating a question framework is to determine the scope of the material and serve as a guide in the preparation of questions. The problem framework will be prepared based on the material or topic to be evaluated, indicators and levels of cognition to be measured, and evaluation methods to be used (Rustanto & Prayitno, 2023).

Grid Arrangement

Creating a problem item framework is a crucial aspect of the design (L. L. Syarifah et al., 2020). The problem framework to be prepared refers to the 2013 or the applicable curriculum. This alignment ensures content validity and curricular relevance of the developed question bank. The purpose of creating a question framework is to determine the scope of the material and serve as a guide in the preparation of questions (Ardellea & Hamdu, 2022). The problem framework will be prepared based on the material or topic to be evaluated, indicators and levels of cognition to be measured, and evaluation methods to be used (Khoerunnisa & Aqwal, 2020). The use of a structured grid also minimizes item bias and supports systematic distribution of cognitive levels across test items.

Practically, the grid used in this study included columns for: (a) content domain, (b) specific learning indicator, (c) cognitive level (based on revised Bloom), (d) item type, and (e) intended difficulty range. This structured approach allowed the research team to monitor the representation of HOTS items and to identify gaps (e.g., underrepresentation of high-level synthesis/evaluation items) before field trials.

Instrument Planning Results

The instrument design stage aims to create an initial framework for collecting assessment instrument data (Destiana et al., 2020). The step is to make question items and validation questionnaire sheets (Utami et al., 2021). In making multiple-choice questions, several aspects must be considered, namely material, construction, and language (Rosdiana et al., 2022). A total of 40 multiple-choice items were developed to represent three core thematic units in the Arabic curriculum, ensuring proportional representation of instructional content.

Table 2. List of Questions in the Arabic Question Bank

Number of Questions	Material
14 Questions	البيانات الشخصية
14 Questions	المرافق العامّة في المدرسة
12 Questions	الحياة في الأسرة وفي سكن الطلاّب

This distribution reflects curricular priorities and ensures content balance across daily communication, school context, and social life themes. In addition, researchers also make answer sheets and instructions for working on questions, which are combined into one question package. Clear instructions are essential in CBT contexts to reduce construct-irrelevant variance caused by misunderstanding test procedures. For scoring guidelines, a dichotomous model was used, which scores 1 for correct and 0 for incorrect answers. This model supports objective scoring and facilitates further item analysis.

Additionally, the digital package included metadata for each item (item code, cognitive level, correct key, and distractor rationales) to support later item banking, rotation, and reporting. In CBT implementation, this metadata enables item randomization while preserving blueprint constraints (Chituc et al., 2019).

Expert and Practitioner Validation Results

After making the answer key sheet, the researcher developed a validation questionnaire sheet used by expert validators and practitioners. This dual validation approach strengthens content and practical validity. The assessed aspects included grids, materials, constructs, and language. At this stage, validators assessed 19 aspects, each scored from 0 to 4. This multi-aspect evaluation provides a comprehensive picture of instrument quality.

Table 3. Validation Test Results by Lecturer Expert Validators

No.	Aspects of Interest	Score
1	Identity congruence	4
2	Suitability of essential curriculum competencies	4
3	Suitability of subject matter with essential competencies	4
4	Alignment of indicators with Basic Competencies and subject matter.	4
5	Question items according to indicators	4
6	There is only one answer key	4
7	The content of the material is the purpose of measurement	4
8	The content of the material is according to the level, type of school, and grade level	4
9	It contains clear instructions on how to solve the problem	4
10	The formulation of the question items does not provide clues towards the correct answer	4
11	Completeness of test instrument contents	4
12	Compatibility between the question items and the characteristics of HOT's questions	1
13	Compatibility between question items and grids	4
14	The subject matter contains one problem that will be stated	4
15	Has the most correct answer key	4
16	The trickster works	3
17	The test instrument uses language that is by the rules of Indonesian	4
18	The question does not contain a double interpretation	4
19	Communicative questions in simple language are easy for students to understand.	4
	Sum	**72**
	Average	**3,79**
	Criteria	**Highly Valid**

The expert validator's average score of 3.79 indicates that the instrument meets very high validity standards, although one aspect related to HOTS characteristics received a lower score, signaling a need for refinement.

Table 4. Validation Test Results by Practitioners

No.	Aspects of Interest	Score
1	Identity field match	4
2	Suitability of applicable curriculum essential competencies	4
3	Suitability of subject matter with essential competencies	4
4	Conformity of the formulation of indicators with the Basic Competencies and subject matter	4
5	Question items according to the question indicator	4
6	There is only one answer key	4
7	The content of the material is the purpose of measurement	4
8	The content of the material is according to the level, type of school, and grade level	3
9	It contains clear instructions on how to solve the problem	4
10	The question items are designed without clues to the correct answer.	3
11	Completeness of test instrument contents	4
12	Compatibility between the question items and the characteristics of HOT's questions	1
13	Compatibility between question items and grids	3
14	The subject matter contains one problem that will be stated	4
15	Has the most correct answer key	3
16	The trickster works	4
17	The test instrument uses language that is by the rules of Indonesian	4
18	The question does not contain a double interpretation	3
19	Communicative questions using simple and easy-to-understand language	4
	Sum	**68**
	Average	**3,58**
	Criteria	**Highly Valid**

Practitioner validation yielded an average score of 3.58, confirming that the instrument is also feasible and appropriate for real classroom implementation.

The pattern of validation scores—strong content alignment but weaker HOTS alignment—suggests that while items accurately reflect curriculum content, the cognitive demand of some items may fall short of intended HOTS targets. This informed the revision strategy: focus on raising item cognitive demand through scenario-based stems, multi-step inference, and improved distractor design, while preserving language appropriateness for Grade X.

Design Revision and Interpretation

Next, the design revision stage is carried out by paying attention to the assessment score and notes of improvement suggestions provided by validators. Interestingly, no major revisions were required at this stage, indicating strong initial design quality. This suggests that the instrument was already aligned with curriculum standards and classroom needs.

Nevertheless, targeted revisions were planned and implemented for items identified with HOTS misalignment and for several items with ambiguous phrasing. These minor but conceptually important changes were aimed at improving cognitive demand and reducing construct-irrelevant variance without altering content coverage.

Product Trial Results and Empirical Evidence

First Trial

The validated test instrument products were then tested for the first stage of test questions on test subjects, namely X-A and X-D class students at State Islamic High School (MAN) 1 Pohuwato. This first trial involved 36 students who were asked to try to do the questions within 40 minutes using the CBT software used in MAN 1 Pohuwato. The first trial focused primarily on item readability, clarity of instructions, and technical usability of the CBT system.

Second Trial

After the product is revised and improved based on the first trial, a second trial phase will be carried out on the same test subjects, namely class X-A and X-D students at State Islamic High School (MAN) 1 Pohuwato, totalling the same 37 students. The explanation of the product revision has been presented in the product revision discussion below. This iterative testing process strengthens empirical validity by ensuring that improvements are evidence-based rather than theoretical.

Question Validity Test

Based on the results of logical validity by experts and practitioners, it can be seen that the number of expert validator scores is 74, with an average score of 3.89 for very valid categories, and the number of scores from practitioner validators is 71, with an average score of 3.74 very valid categories:

Table 5. Results of Logical Validity by Experts and Practitioners

Validator	Sum Score	Average Score	Category
Expert Validators	72	3,79	Highly Valid
Practitioner Validator	68	3,58	Highly Valid
Average Number of Scores		7,63	
Number of Validators		2	
The average of both validators		3,68	Highly Valid

The average result of validity by both validators is 3.815, so the design or problem design developed by the researcher is said to be very valid. After carrying out validation tests by experts and practitioners, researchers conducted question trials in 2 classes, namely classes X-A and X-D State Islamic High School (MAN) 1 Pohuwato, with a total of 36 students. Based on the results of these trials, the empirical validity of each question item developed will be known. Test the validity of the question items using biserial correlation, where an item is declared valid when its point-biserial correlation exceeds the critical value (r = 0.316) at α = 0.05). To provide a clearer picture of item-level validity patterns across the test, the distribution of point-biserial correlations is visualized in Figure 2."

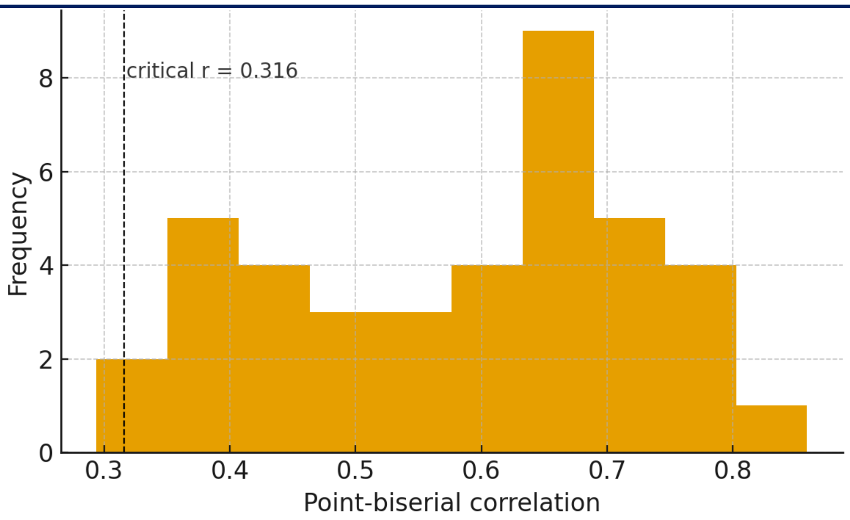

Figure 2. Distribution of Point-Biserial Correlations

“As shown in Figure 4, 39 out of 40 items exceed the critical point-biserial threshold (r = 0.316), supporting the strong empirical validity of the developed question bank

a. Empirical Validity Results

Table 6. Empirical Validity Analysis Table of Question Items

Question Number	Calculated value	Stable Value	Category
1	0,769	0,3160	Valid
2	0,859	0,3160	Valid
3	0,394	0,3160	Valid
4	0,728	0,3160	Valid
5	0,463	0,3160	Valid
6	0,728	0,3160	Valid
7	0,317	0,3160	Valid
8	0,762	0,3160	Valid
9	0,392	0,3160	Valid
10	0,705	0,3160	Valid
11	0,294	0,3160	Invalid
12	0,762	0,3160	Valid
13	0,472	0,3160	Valid
14	0,519	0,3160	Valid
15	0,688	0,3160	Valid
16	0,457	0,3160	Valid
17	0,624	0,3160	Valid
18	0,558	0,3160	Valid
19	0,688	0,3160	Valid
20	0,733	0,3160	Valid

21	0,630	0,3160	Valid
22	0,660	0,3160	Valid
23	0,558	0,3160	Valid
24	0,688	0,3160	Valid
25	0,476	0,3160	Valid
26	0,688	0,3160	Valid
27	0,414	0,3160	Valid
28	0,383	0,3160	Valid
29	0,374	0,3160	Valid
30	0,684	0,3160	Valid
31	0,688	0,3160	Valid
32	0,775	0,3160	Valid
33	0,688	0,3160	Valid
34	0,428	0,3160	Valid
35	0,618	0,3160	Valid
36	0,383	0,3160	Valid
37	0,688	0,3160	Valid
38	0,607	0,3160	Valid
39	0,545	0,3160	Valid
40	0,733	0,3160	Valid
	0,591	**0,3160.**	

Based on Table 5, 39 out of 40 items are empirically valid. The presence of one invalid item reflects normal variation in test development and does not undermine the overall instrument quality. Question number 11 was invalid due to low biserial correlation, indicating inconsistency between item performance and total score. This finding underscores the importance of empirical testing even after expert validation.

Further investigation of Item 11's response pattern suggested possible causes: ambiguous wording, mis-keying, or that the item measured a slightly different sub-skill than the rest of the bank. Post-hoc item review recommended rewriting the stem and distractors and re-testing in a subsequent field administration. This iterative empirical correction exemplifies best-practice test development.

Reliability Analysis

Reliability test based on the results of field trials on test subjects as many as 36 students from classes X-A and X-B State Islamic High School (MAN) 1 Pohuwato. To determine the reliability test, researchers used the help of the SPSS version 26 application for Windows. The analysis results of the overall reliability test calculation of test questions will be described below.

Table 7. Reliability Test Results of the developed question items

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,952	**,952**	**40**

A Cronbach's Alpha of 0.952 indicates very high internal consistency, exceeding commonly accepted reliability thresholds. This result demonstrates that the instrument provides stable and consistent measurements.

To complement the reliability claim, it is recommended to report the 95% confidence interval for Cronbach's alpha and the "Cronbach's alpha if item deleted" table in the final manuscript; these additional statistics inform whether any single item adversely affects internal consistency.

The high validity scores obtained from expert validators (3.79) and practitioner validators (3.58), as well as the very high reliability coefficient (Cronbach's Alpha = 0.952), can be attributed to several methodological factors in the development process. First, the construction of the test items was guided by a detailed test blueprint that aligned each item with specific curriculum competencies and learning indicators. Second, the items were systematically reviewed by both assessment experts and experienced Arabic teachers, ensuring conceptual accuracy, linguistic clarity, and instructional relevance. Third, iterative revisions based on validator feedback helped eliminate ambiguous wording and weak distractors, thereby strengthening item consistency. These procedures are widely recognized in educational measurement literature as key contributors to high instrument validity and reliability.

Difficulty Level Analysis

Researchers analyzed the difficulty level in the developed question bank with the help of the SPSS program version 26 for Windows, and the difficulty level results came from the *product trials,* which will be presented in detail in the table below.

Table 8. Results of Question Difficulty Analysis

Question Number	Difficulty Level	Category	Question Number	Difficulty Level	Category
1	0,81	Easy	21	0,70	Easy
2	0,78	Easy	22	0,81	Easy
3	0,78	Easy	23	0,81	Easy
4	0,81	Easy	24	0,70	Easy
5	0,76	Easy	25	0,26	Difficult
6	0,81	Easy	26	0,70	Easy
7	0,78	Easy	27	0,76	Easy
8	0,78	Easy	28	0,76	Easy
9	0,73	Easy	29	0,73	Easy
10	0,76	Easy	30	0,59	Keep
11	0,81	Easy	31	0,70	Easy

12	0,28	Difficult	32	0,78	Easy
13	0,68	Keep	33	0,30	Difficult
14	0,86	Easy	34	0,65	Keep
15	0,70	Easy	35	0,70	Easy
16	0,86	Easy	36	0,76	Easy
17	0,84	Easy	37	0,70	Easy
18	0,81	Easy	38	0,57	Keep
19	0,70	Easy	39	0,65	Keep
20	0,68	Keep	40	0,68	Keep
Easy Questions		**30 Question**			
Medium Problem		**7 Question**			
Difficult Problems		**3 Question**			
Number of Questions		**40 Question**			

Based on the results of the analysis of the difficulty level of the question items, it can be seen that 30 questions fall into the easy category, namely numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 26, 27, 28, 29, 31, 32, 35, 36 and 37 which means that many students answer these questions correctly. Seven questions fall into the medium difficulty category, namely numbers 13, 20, 30, 34, 38, 39, and 40, which means that students who answer the right questions are balanced with students who answer incorrectly. Then, 3 questions fall into the problematic category, namely number 12, 25, and question number 33, only because very few students can answer the questions correctly.

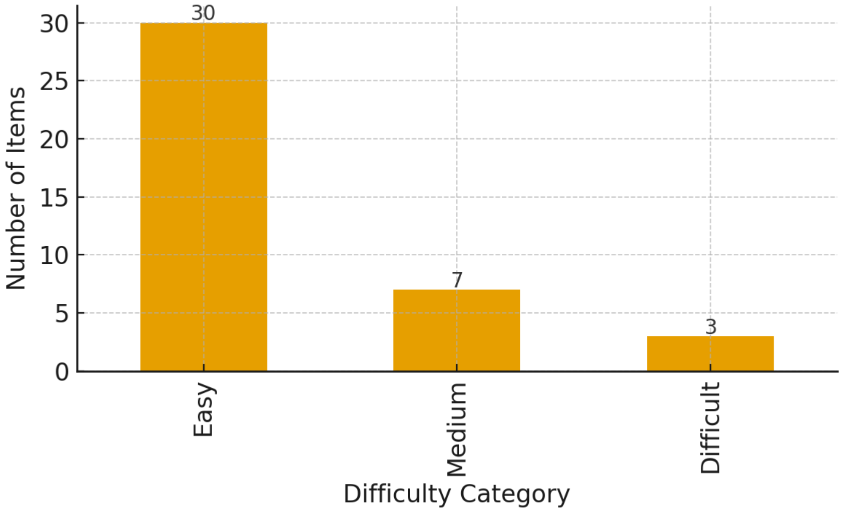

Figure 3. Distribution of Difficulty Categories

The predominance of easy items indicates that the instrument is well-suited for diagnosing basic mastery and encouraging learner success in formative settings. However, for high-stakes summative assessments or for better measurement of higher-order thinking, the bank should be expanded with more medium and difficult items that require analysis, synthesis, and evaluation. A deliberate item-writing workshop focused on HOTS item design is recommended for future iterations.

Test the Discriminating Power of the Question

Question differentiating power is a condition where a question can distinguish between high-ability and low-ability students. The question is terrible for question items that can be answered correctly by students with high and low abilities because it has no discriminating power, and vice versa if students with high and low skills cannot answer the questions correctly. Good questions can be answered correctly by students with high skills. The attached data shows an analysis of differentiating power through the SPSS application version 26 for Windows—the following results are based on analyzing the distinguishing power of question items in the Arabic subject question bank.

Table 9. Results of the Differentiating Power Analysis of Question Points

Question Number	Result in SPSS	Interpretation	Question Number	Result in SPSS	Interpretation
S01	0,752	Very Good	**S21**	0,602	Good
S02	0,848	Very Good	**S22**	0,638	Good
S03	0,359	Enough	**S23**	0,531	Good
S04	0,709	Excellent	**S24**	0,663	Good
S05	0,429	Good	**S25**	0,442	Good
S06	0,709	Excellent	**S26**	0,663	Good
S07	0,280	Enough	**S27**	0,378	Enough
S08	0,744	Excellent	**S28**	0,346	Enough
S09	0,354	Enough	**S29**	0,336	Enough
S10	0,683	Good	**S30**	0,657	Good
S11	0,258	Enough	**S31**	0,663	Good
S12	0,744	Excellent	**S32**	0,758	Excellent
S13	0,435	Good	**S33**	0,663	Good
S14	0,494	Good	**S34**	0,388	Enough
S15	0,663	Good	**S35**	0,589	Good
S16	0,430	Good	**S36**	0,346	Enough
S17	0,601	Good	**S37**	0,663	Good
S18	0,531	Good	**S38**	0,575	Good
S19	0,663	Good	**S39**	0,510	Good
S20	0,710	Excellent	**S40**	0,710	Excellent

Differentiating Power Category	**Number of Questions**
Excellent	9

Good	22
Enough	9
Signs	0
Very ugly	0

Based on table 8, it is known that questions number 1, 2, 4, 6, 8, 12, 20, 32, and 40 have a distinguishing power with the category "outstanding," meaning that these questions are perfect for distinguishing students with high ability from students with low ability. Questions number 5, 10, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 30, 31, 33, 35, 37, 38 and 39 are in the "good" category, meaning that they are suitable for distinguishing high-ability students from low-ability students.

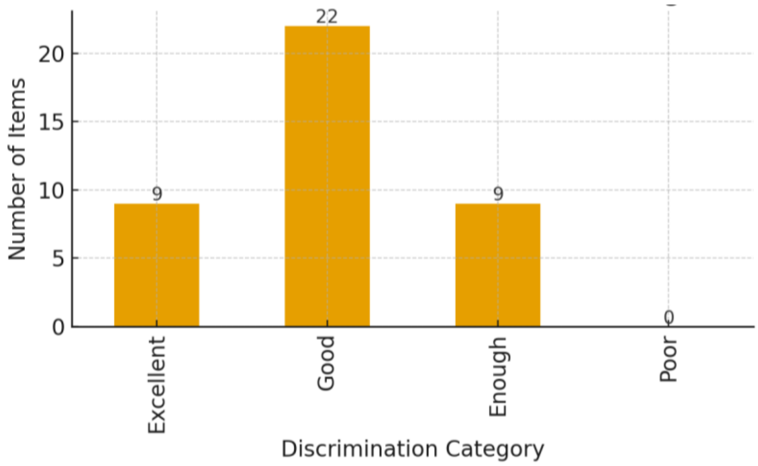

Figure 4. Distribution of Item Discrimination Categories

Next are questions number 3, 7, 9, 11, 27, 28, 29, 34, and 36, which have distinguishing power with the category "enough," meaning that the questions made are enough to determine students who are high in ability from students who are low in ability. However, there still needs to be a revision process to see if it can be improved. The last is in the category of "ugly" and very ugly, distinguishing power amounting to 0 questions or none. Questions are said to be ugly or very ugly because these questions are answered more by students with low abilities. Things like this can happen because the deceiver that has been made usually does not work correctly. Most items fell into good to excellent categories, indicating strong discriminatory capacity. Items with sufficient discriminating power may be revised to improve distractor effectiveness and cognitive demand.

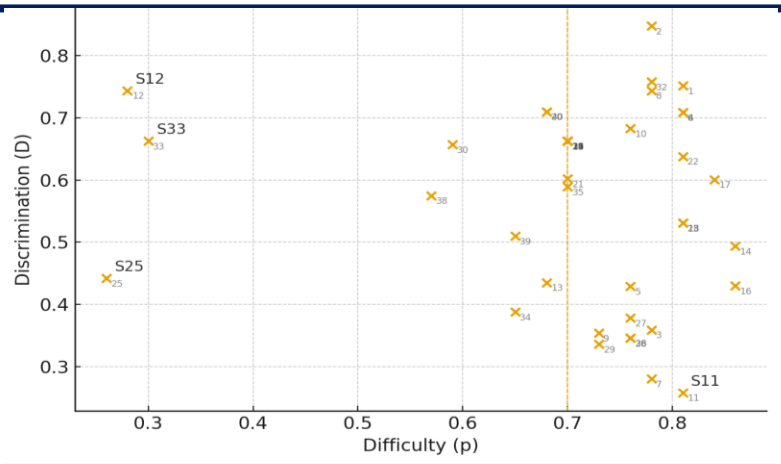

Figure 5. Item Difficulty VS Discrimination

Figure 5 plots item difficulty against discrimination; flagged items (e.g., Item 11, 12, 25, 33) are highlighted as candidates for revision."

Most items show good to excellent discrimination, which is a robust indicator of item quality. Items with only sufficient discrimination were reviewed and revised (see Table 9) to improve distractor plausibility and alignment with the targeted cognitive skill. No items showed negative discrimination, which indicates no items were systematically favoring lower-ability students.

Product Revision Outcomes

Based on the results of the first trial that has been described earlier, it produces a revision of the product of question items that need to be improved. The following are the question items that need to be improved, namely questions number 3, 6, 8, 13, 16, 17, 25, 28, 30, 31 and 39.

Table 10. Question Revision Table

Number of Question	Before Revision	After Revision
3	عَالِيٌّ : السَّلَامُ عَلَيْكُمْ يَاعُمَرُ, كَيْفَ حَالُكَ؟ عُمَرُ : وَعَلَيْكُمُ السَّلَامُ, أَنَا بِخَيْرٍ وَالحَمْدُ لِلَّهِ. وَ أَنْتَ؟ عَالِيٌّ : بِخَيْرٍ وَالحَمْدُ لِلَّهِ, مَااسْمُكَ بِالكَامِلِ؟ عُمَرُ : .... عَالِيٌّ : طَيِب.	عَالِيٌّ : السَّلَامُ عَلَيْكُمْ يَاعُمَرُ, كَيْفَ حَالُكَ؟ عُمَرُ : وَعَلَيْكُمُ السَّلَامُ, أَنَا بِخَيْرٍ وَالحَمْدُ لِلَّهِ. وَ أَنْتَ؟ عَالِيٌّ : بِخَيْرٍ وَالحَمْدُ لِلَّهِ, مَااسْمُكَ بِالكَامِلِ؟ عُمَرُ : .... عَالِيٌّ : طَيِب. What are the right conversational sentences to fill in the above dots?
6	The meaning of the sentence underlined in the Qiro'ah text above is.....	وَعُنْوَانُ المَدْرَسَةِ: شَارِعُ اَحْمَدْ دَحْلَانُ. What is the exact translation for the following sentence?

8		 Which Arabic sentence, as per the picture above, is?
13	فِي مَدْرَسَتِنَا مَرَ افِقَ عَامَّةً كَثِيرَةً	فِي مَدْرَسَتِنَا مَرَ افِقَ عَامَّةً كَثِيرَةً The exact translation of the sentence snippet underlined in the Qira'ah text above is.....
16		 Is the right Arabic sentence to explain the picture above?
17	أَقُوْمُ مِنَ النَّوْمِ فِي السَّاعَةِ الرَّابِعَةِ	أَقُوْمُ مِنَ النَّوْمِ فِي السَّاعَةِ الرَّابِعَةِ The meaning of the underlined word in the sentence above is?
25	أَنَا – أَنَا – الصَّفِّ – عَبْدُ الرَّزَقِ – طَالِبٌ – الأَوَّلِ – مِنْ- طُلَّابٌ	أَنَا – أَنَا – الصَّفِّ – عَبْدُ الرَّزَقِ – طَالِبٌ – الأَوَّلِ – مِنْ- طُلَّابٌ The exact wording of the above sentence is?
28	بَعْضَ – فِي المَدْرَسَةِ ! – أُذْكُرْ – المَرَ افِقِ العَامَّةِ	بَعْضَ – فِي المَدْرَسَةِ ! – أُذْكُرْ – المَرَ افِقِ العَامَّةِ The exact wording of the above sentence is?
30	نَجْتَمِعُ – فِي – نَحْنُ – فِي – المَسَاءِ – غُرْفَةُ – الأُسْرَةِ.	نَجْتَمِعُ – فِي – نَحْنُ – فِي – المَسَاءِ – غُرْفَةُ – الأُسْرَةِ. The exact wording of the above sentence is?
31	وَهذِهِ خَيْرُ النِّسْوَةِ. هِيَ طَالِبَةٌ مَاهِرَةٌ	وَهذِهِ خَيْرُ النِّسْوَةِ. هِيَ طَالِبَةٌ مَاهِرَةٌ The meaning of the underlined word in the sentence above is?

39		 Is the right Arabic sentence to explain the picture above?

In the revision of the questions above, most revisions are the instructions for filling in the questions on each question item and the explanation or order of the illustrated questions. After completing the revision, students are expected to understand the question instructions well before answering them. The results of the revised question bank can be seen on the appendix page.

The findings demonstrate that the developed Arabic question bank is valid, reliable, and suitable for CBT-based assessment. The results contribute to the growing body of research on digital assessment development in language education.

However, this study is limited by its relatively small sample size and single institutional context. Future studies should involve larger and more diverse samples and explore the impact of the question bank on learning outcomes over time.

Additional note: post-revision, it is important to re-run empirical analyses (validity, difficulty, discrimination) on the revised items to document whether changes improved psychometric properties. A short re-test or equivalence study is recommended for items that were substantially altered.

The findings demonstrate that the developed Arabic question bank is valid, reliable, and suitable for CBT-based assessment. This supports the central hypothesis of the study and provides empirical evidence that a curriculum-aligned, HOTS-aware multiple-choice bank can be implemented effectively within a madrasah context.

Practically, the instrument offers teachers a valid and reliable tool for UTS/UAS that reduces subjective scoring and administrative burden. Technologically, the CBT implementation improved scoring efficiency and test security through randomization and automated marking. Pedagogically, the instrument can guide instruction by identifying topic areas needing reinforcement (using item-level statistics).

The findings contribute to the growing body of research on digital assessment development in language education. However, this study is limited by its relatively small sample size and single institutional context. Future studies should involve larger and more diverse samples and explore the impact of the question bank on learning outcomes over time.

Additional specific recommendations:

1) Expand the item bank to include more medium and difficult items, and to incorporate constructed-response items for productive skills (writing and speaking).
2) Collect formal usability data (system logs, time-on-item, and standardized usability questionnaires) to quantify CBT performance and student experience.

3) Report additional psychometric details in the final manuscript: 95% CI for Cronbach's alpha, Cronbach-if-item-deleted, p-values for point-biserial correlations, and a short DIF analysis if sample size permits.

DISCUSSION

The results of this study indicate that the Arabic question bank developed through a systematic Research and Development (R&D) process demonstrates strong psychometric quality. This is evidenced by very high validity scores from expert validators (3.79) and practitioner validators (3.58), as well as excellent internal consistency (Cronbach's Alpha = 0.952). These findings are in line with research in educational measurement, which emphasizes that assessment instruments developed through expert judgment, blueprint alignment, and iterative revision tend to produce higher levels of validity and reliability (Hannon et al., 2019). In language assessment contexts, careful alignment between test items and curriculum competencies is essential to ensure that assessment results accurately reflect learners' actual proficiency (Wallace & Ke, 2023).

Further analysis showed that 39 out of 40 items were statistically valid, with most items demonstrating acceptable to excellent levels of difficulty and discriminating power. This result supports previous findings that well-constructed multiple-choice items can function effectively in both paper-based and computer-based assessments, particularly when distractors are designed to reflect common learner misconceptions (Eleragi et al., 2025). Nevertheless, the tendency toward a higher proportion of easy items identified in this study echoes concerns raised by earlier research, which notes that assessment instruments often underrepresent higher-order thinking skills unless HOTS principles are intentionally embedded during item construction (Zhang, 2021).

A key contribution of this study lies in its integrative approach, combining the development of a validated Arabic question bank with the implementation of a CBT system in a madrasah context. While previous studies have examined CBT adoption and test construction separately, this research demonstrates that digital assessment platforms must be supported by psychometrically sound instruments to ensure meaningful and trustworthy evaluation outcomes (Cáceres et al., 2020). In the context of Arabic language education, where assessment practices remain largely conventional, this integration offers a practical and replicable model for aligning curriculum standards, HOTS-oriented assessment, and digital testing systems.

From a pedagogical and practical perspective, the developed question bank provides Arabic teachers with an efficient and objective tool for administering the Mid-Term Examination (UTS) and Final Examination (UAS). CBT-based assessment reduces scoring subjectivity, improves administrative efficiency, and enhances test security, as consistently reported in recent studies on digital assessment in secondary education (Niimi & Matsuura, 2022). However, this study is limited by its implementation in a single institution with a relatively small sample size, which may restrict the generalizability of the findings. Future research is therefore recommended to apply the instrument in broader educational contexts, refine item difficulty distribution to better capture higher-order cognitive processes, and examine the long-term impact of CBT-based assessment on students' learning outcomes and instructional improvement.

CONCLUSION

This development research produced an Arabic question bank for Grade X students at State Islamic High School (MAN) 1 Pohuwato through three systematic stages—define, design, and develop—ensuring methodological rigor, content relevance, and empirical soundness. The product consists of 40 multiple-choice items representing three core thematic units in the Grade X Arabic curriculum and was designed to be applicable in both conventional paper-based tests and CBT environments. Validation results showed very high validity, with average scores of 3.79 from expert validators and 3.58 from practitioner validators, while empirical testing indicated that 39 out of 40 items were valid and the instrument demonstrated very high reliability (Cronbach's Alpha = 0.952), along with acceptable to excellent levels of difficulty and discriminating power. These findings confirm that the developed Arabic question bank is technically sound, pedagogically appropriate, and practically feasible for use in Mid-Term Examination and Final Examination, supporting objective, efficient, and HOTS-oriented evaluation practices aligned with digital assessment demands in secondary education. Although the study was limited to one institution and a relatively small sample size, the results provide an empirically validated model for integrating curriculum standards, higher-order thinking skills, and CBT-based assessment, while future research is recommended to expand implementation scope, refine item difficulty distribution, and examine the long-term impact of CBT-based assessment on students' learning outcomes.

REFERENCES

Alkhatib, O. J. (2022). An Effective Assessment Method of Higher-Order Thinking Skills (Problem-Solving, Critical Thinking, Creative Thinking, and Decision-Making) in Engineering and Humanities. *2022 Advances in Science and Engineering Technology International Conferences, ASET 2022*. https://doi.org/10.1109/ASET53988.2022.9734856

Ardellea, F., & Hamdu, G. (2022). Pentingnya Kemampuan Guru Sekolah Dasar dalam Mengembangkan Soal Tes Literasi dan Numerasi Berbasis Education for Sustainable Development (ESD). *Edu Cendikia: Jurnal Ilmiah Kependidikan*, *2*(02), 220–227. https://doi.org/10.47709/educendikia.v2i02.1587

Bahrun, Maulana, R., Muslem, A., & Yulianti. (2023). Designing Assessment, Learning Strategies, and Obstacles in Facing Computer-Based Madrasah Exam on the English Subject. *Studies in English Language and Education*, *10*(2), 884 – 906. https://doi.org/10.24815/siele.v10i2.31954

Beerepoot, M. T. P. (2023). Formative and Summative Automated Assessment with Multiple-Choice Question Banks. *Journal of Chemical Education*, *100*(8), 2947–2955. https://doi.org/10.1021/acs.jchemed.3c00120

Cáceres, M., Nussbaum, M., & Ortiz, J. (2020). Integrating critical thinking into the classroom: A teacher's perspective. *Thinking Skills and Creativity*, *37*, 100674. https://doi.org/https://doi.org/10.1016/j.tsc.2020.100674

Chituc, C.-M., Herrmann, M., Schiffner, D., & Rittberger, M. (2019). Towards the Design and Deployment of an Item Bank: An Analysis of the Requirements Elicited. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11841 LNCS*, 155–162. https://doi.org/10.1007/978-3-030-35758-0_15

Desiriah, E., & Setyarsih, W. (2021). Tinjauan Literatur Pengembangan Instrumen Penilaian Kemampuan Berpikir Tingkat Tinggi (Hots) Fisika Di Sma. *ORBITA: Jurnal Kajian, Inovasi Dan Aplikasi Pendidikan Fisika*, *7*(1), 79. https://doi.org/10.31764/orbita.v7i1.4436

Destiana, D., Suchyadi, Y., & Anjaswuri, F. (2020). Pengembangan instrumen penilaian untuk meningkatkan kualitas pembelajaran produktif di sekolah dasar. *Jurnal Pendidikan Dan Pengajaran Guru Sekolah Dasar (JPPGuseda)*, *3*(2), 119–123. https://doi.org/https://doi.org/10.55215/jppguseda.v3i2.2720

Eleragi, A. M. S., Miskeen, E., Hussein, K., Rezigalla, A. A., Adam, M. I. E., Al-Faifi, J. A., Alhalafi, A., Al Ameer, A. Y., & Mohammed, O. A. (2025). Evaluating the multiple-choice questions quality at the College of Medicine, University of Bisha, Saudi Arabia: a three-year experience. *BMC Medical Education*, *25*(1). https://doi.org/10.1186/s12909-025-06700-2

Gebril, A. (2017). Arabic language teachers' conceptions of assessment and the hidden tension between accountability and improvement in Egyptian schools. In *The Routledge Handbook of Arabic Linguistics* (pp. 560–574). https://doi.org/10.4324/9781315147062

Hakami, Y. A. A., Hussin, A. R. B. C., & Dahlan, H. M. (2015). Technology acceptance for CBT in secondary schools of Saudi Arabia. *Proceedings - International Conference on Intelligent Systems, Modelling and Simulation, ISMS*, *2015-September*, 804–807. https://doi.org/10.1109/ISMS.2014.148

Hannon, P., Lappe, K., Griffin, C., Roussel, D., Colbert-getz, J., Roussel, D., & Colbert-getz, J. (2019). An objective structured clinical examination : From examination room to Zoom breakout room. *Medical Education Adaptations*, *54*(9), 14241. https://doi.org/10.1111/medu.14241

Hastini, L. Y., Fahmi, R., & Lukito, H. (2020). Apakah Pembelajaran Menggunakan Teknologi dapat Meningkatkan Literasi Manusia pada Generasi Z di Indonesia? *Jurnal Manajemen Informatika (JAMIKA)*, *10*(1), 12–28. https://doi.org/10.34010/jamika.v10i1.2678

Hatamnejad, M. R., Shirvani, M., Pourhoseingholi, M. A., Balaii, H., Shahrokh, S., Aghdaei, H. A., Koolaeian, A., & Cheraghpour, M. (2023). Translation and cross-cultural adaptation of the Persian version of inflammatory bowel disease-fatigue (IBD-F) self-assessment questionnaire. *PLoS ONE*, *18*(7 July). https://doi.org/10.1371/journal.pone.0288592

Khoerunnisa, P., & Aqwal, S. M. (2020). Analisis Model-model Pembelajaran. *Fondatia*, *4*(1), 1–27. https://doi.org/10.36088/fondatia.v4i1.441

Krzic, M., & Brown, S. (2022). Question banks for effective online assessments in introductory science courses. *Natural Sciences Education*, *51*(2). https://doi.org/10.1002/nse2.20091

Luecht, R. M. (2015). Computer-Based Test Delivery Models, Data, and Operational Implementation Issues. In *Technology and Testing: Improving Educational and Psychological Measurement* (pp. 179–205). https://doi.org/10.4324/9781315871493-10

Makruf, I., & Barokah, A. (2023). Improving the Quality of ICT-Based Arabic Learning Assessment With Online Applications. *Journal of Higher Education Theory and Practice*, *23*(11), 32 – 40. https://doi.org/10.33423/jhetp.v23i11.6216

Malik, R. H., & Malik, A. S. (2020). Developing A Bank of Faculty-Authored, Valid and Reliable Objective Questions for Institutional Use: Sharing the Experience. *Malaysian Journal of Medicine and Health Sciences*, *16*, 28–35. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092237321&partnerID=40&md5=b5fb7514878af72131c6e6e43c2332f5

Moh Firdaus, Hula, I. R. N., & Bahri, R. B. H. (2023). *Arabic Language Teacher Website Media Development in the Teaching and Learning Process of Qiraah Material*. *2*(3), 353–366. https://doi.org/https://doi.org/10.58194/eloquence.v2i3.1424

Mohamad Aso Samsudin, & Ukhtul Iffah. (2019). Penilaian Autentik Pada Matapelajaran Pendidikan Agama Islam. *Edupedia*, *4*(1), 77–85. https://doi.org/10.35316/edupedia.v4i1.528

Mujiatun, S., Jasin, H., Fahmi, M., & Jufrizen, J. (2022). Model Financial Technology (Fintech) Syariah di Sumatera Utara. *Owner*, *6*(3), 1709–1718. https://doi.org/10.33395/owner.v6i3.893

Muttaqin, B. I. A. (2019). Telaah Kajian dan Literature Review Design of Experiment (DoE). *Journal of Advances in Information and Industrial Technology*, *1*(1), 33–40. https://doi.org/10.52435/jaiit.v1i1.10

Natadireja, U., & Nurachadijat, K. (2023). Evaluasi Pendidikan Menuju Insan Kamil dalam Perspektif Filsafat Ilmu. *Al-Idaroh: Jurnal Studi Manajemen Pendidikan Islam*, *7*(2), 253–267. https://doi.org/10.54437/alidaroh.v7i2.929

Niimi, N., & Matsuura, N. (2022). Assessing Japanese junior high school students' English achievement through computer-based testing in the classroom: a case of integrated reading-into-writing continuous task. *Language Testing in Asia*, *12*(1). https://doi.org/10.1186/s40468-022-00189-y

Noval, Y., & Adhani, A. (2021). Analisis Soal Penilaian Akhir Semester Pada Mata Pelajaran Biologi Kelas X Berdasarkan Taksonomi Anderson Di Sma Negeri 1 Tarakan. *Borneo Journal Of Biology Education*, *3*(1), 18–28. https://doi.org/https://doi.org/10.35334/bjbe.v3i1.1887

Nugroho, H. Y. S. H., Basuki, T. M., Pramono, I. B., Savitri, E., Purwanto, Indrawati, D. R., Wahyuningrum, N., Adi, R. N., Indrajaya, Y., Supangat, A. B., Putra, P. B., Auliyani, D., Priyanto, E., Yuwati, T. W., Pratiwi, Narendra, B. H., Sukmana, A., Handayani, W., Setiawan, O., & Nandini, R. (2022). Forty Years of Soil and Water Conservation Policy, Implementation, Research and Development in Indonesia: A Review. *Sustainability (Switzerland)*, *14*(5). https://doi.org/10.3390/su14052972

Pellas, N. (2025). Exploring learning outcomes and psycho-emotional experiences of undergraduate students in digital literacy training and support using Web-based assessment platforms. *Journal of Research on Technology in Education*, *57*(4), 820–841. https://doi.org/10.1080/15391523.2024.2323451

Punoševac, M., & Nikolić, A. (2024). AI in Action: Transforming Assessment Practices – A Case Study of YouTestMe. *CEUR Workshop Proceedings*, *3938*, 139–151. https://www.scopus.com/inward/record.uri?eid=2-s2.0-105000232374&partnerID=40&md5=bbd5029f455a0cfb64c165b05e1bc18c

Rayón, A., Guenaga, M., & Núñez, A. (2014). Integrating and visualizing learner and social data to elicit higher-order indicators in SCALA dashboard. *ACM International Conference Proceeding Series*, *16-19-September-2014*. https://doi.org/10.1145/2637748.2638435

Reza Amirian, S. M., Abbasi, F. M., & Zolfagharkhani, M. (2023). EFL Learners' Perspective towards Online Assessments during COVID-19 Outbreak. *International Journal of Language Testing*, *13*(2), 188–205. https://doi.org/10.22034/IJLT.2023.382556.1226

Rosdiana, R., Budiana, S., Mahajani, T., & Talitha, S. (2022). Penerapan HOTS pada Soal-soal Buku Teks Pelajaran Bahasa Indonesia Tingkat SMA. *Aksara: Jurnal Ilmu Pendidikan Nonformal*, *8*(2), 1065. https://doi.org/10.37905/aksara.8.2.1065-1074.2022

Rustanto, P. C. R., & Prayitno, B. A. (2023). Developing Complex Multiple-Choice Test to Empower Students Higher Order Thinking Skill about Excression System. *AIP Conference Proceedings*, *2540*. https://doi.org/10.1063/5.0107968

Strother, J. B., Fazal, Z., Johnson, A., & Millsap, M. (2008). Effects of test modality - Computer-based versus Paper-and-Pencil - On nonnatives' English results. *Proceedings of the 7th IASTED International Conference on Web-Based Education, WBE 2008*, 133–138. https://www.scopus.com/inward/record.uri?eid=2-s2.0-62949209289&partnerID=40&md5=9bb8a0a1e843a3402ce160f2fbcd2912

Syarifah, L. L., Yenni, & Dewi, W. K. (2020). Analisis Soal-Soal Pada Buku Ajar Matematika Siswa. *Jurnal Cendekia: Jurnal Pendidikan Matematika*, *04*(02), 1259–1272. https://doi.org/https://doi.org/10.31004/cendekia.v4i2.335

Syarifah, N. Y., & Asda, P. (2021). Computer Based Test Sebagai Alternatif Metode Penilaian Evaluasi Pembelajaran Manajemen Keperawatan. *Jurnal Kesehatan Masyarakat*, *14*(2), 478–486. https://doi.org/https://doi.org/10.47317/jkm.v14i2.367

Utami, T. P., Sjaifuddin, S., & Berlian, L. (2021). Pengembangan Soal Uraian Berbasis Indikator Kemampuan Berpikir Tingkat Tinggi pada Konsep Sistem Pencernaan pada Manusia untuk Siswa Kelas VIII SMP/Mts. *PENDIPA Journal of Science Education*, *6*(1), 128–134. https://doi.org/10.33369/pendipa.6.1.128-134

Utomo, I. F. (2021). Pengembangan Media Pembelajaran Video Tutorial Menjahit Lengan Tulip Siswa Kelas X Tata Busana Di SMK Muhammadiyah 1 Tempel. *Jurnal Fesyen: Pendidikan Dan Teknologi*, *10*(1). https://doi.org/https://doi.org/10.21831/teknik%20busana.v10i1.17232

Wallace, M. P., & Ke, H. (2023). Examining the Content Alignment Between Language Curriculum and A Language Test in China. *Teflin Journal*, *34*(1), 116–135. https://doi.org/10.15639/teflinjournal.v34i1/116-135

Yu, W., & Iwashita, N. (2021). Comparison of test performance on paper-based testing (PBT) and computer-based testing (CBT) by English-majored undergraduate students in China. *Language Testing in Asia*, *11*(1). https://doi.org/10.1186/s40468-021-00147-0

Zhang, X. (2021). Preparing first-year college students' academic transition: What is the value of complementary web-based learning? *Computers & Education*, *172*, 104265. https://doi.org/https://doi.org/10.1016/j.compedu.2021.104265