Difficulty Level Of Arabic Language Exam Questions For High School Student

Received 2025-06-23
Accepted 2025-10-20
Published 2026-01-11

Luqi Fauziyah*[1], Nabilah Robbaniyah[2], Aliwafa[3], Moh. Ansori[4], Nila Sari Syafitri[5], Nabiela Fahma[6]
[1,3,4,5,6]Arabic Language Education, UIN Sunan Ampel Surabaya, Indonesia
[2]Universitas Telkom Surabaya, Indonesia
luqifauziyah24@gmail.com*[1], nabilahrobbaniyah@gmail.com[2], aliw87870@gmail.com[3], m.anshori@uinsa.ac.id[4], nillfitri123@gmail.com[5], nabielafahma@gmail.com[6]

To cite this article: Fauziyah, Luqi., Robbaniyah, Nabilah., Aliwafa, Aliwafa., Ansori, Moh., Syafitri, Nila Sari., Fahma, Nabiela. (2026). Difficulty Level Of Arabic Language Exam Questions For High School Student. Ijaz Arabi: Journal of Arabic Learning, 9 (1), 408-420, DOI: https://doi.org/10.18860 /ijazarabi.V9i1.36740

Abstract

Good learning outcome evaluation questions must have a degree of difficulty that matches what is being measured. To obtain good questions, a study has been conducted to assess the degree of difficulty of the test questions. The object of this research is the MA Islamiyah Senori Final Semester Exam questions. The analysis was performed using quantitative methods and a simple descriptive qualitative approach. Data collection was carried out by assessing each exam question. Based on the study's results, the difficulty of each question was evaluated. From the results, it was observed that the exam questions were distributed as follows: 2 difficult, 9 medium, and 9 easy out of 20. It can be inferred from the study's findings that promising findings can be used. While questions categorized as not good indicate that the level of difficulty is unbalanced, they should be used again; however, it is necessary to carefully consider the analysis results of the questions in terms of their level of difficulty.

Keywords: Level of Difficulty; Evaluation; Exam Questions; Analysis; Easy

INTRODUCTION

Assessment is an integral component of the curriculum, as it not only measures students' acquisition of knowledge and skills but also serves as a powerful tool to drive student learning and behavior. In modern education, the model of learning assessment has shifted from traditional assessment to a more contextual and authentic approach. Assessment becomes a crucial element that plays a role not only as a tool for measuring academic achievement, but also as a means of encouraging and improving the learning process.(L. Fauziyah et al., 2025) It is also important to conduct periodic reviews and assessments.(Jundi et al., 2024) Participatory assessment enhances participants' thinking and evaluation skills.(Baihaqi et al., 2025). For educators, evaluation is important as a means of control, assurance, and quality determination, as well as a form of responsibility for education providers.(Dianova & Anwar, 2024) In other words, assessments can influence what students learn and how they learn, a phenomenon known as the "washback" or "driving effect" of assessment. To ensure that assessments serve their purpose effectively, it is essential that assessment items are closely aligned with the curriculum objectives and desired learning outcomes. This alignment ensures that the exam measures what is being taught and what students are expected to learn, thus providing an accurate picture of their competencies. Misalignment between assessments

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

and the curriculum can lead to several problems: students may misdirect their learning efforts, instructors may fail to highlight relevant material, and the overall educational process may become less effective. Therefore, it is important to regularly evaluate assessment items to ensure that they are aligned with the stated curriculum objectives. This will ultimately lead to more accurate measurements of student learning outcomes and better educational outcomes.(Khan et al., 2025)

Testing is one way to evaluate students' achievement and mastery of skills. Through tests, teachers can find out students' achievements and identify their strengths and weaknesses. Therefore, the test items must be of high quality so that they can reflect the achievement of educational goals. The success of Arabic language learning is influenced by teachers, students, and learning objectives. Teachers must be competent in creating and assessing inquiries. Assessment is very important to measure the extent to which learning objectives are achieved and to provide feedback for improving the teaching system. Its standards and quality have not been analyzed thoroughly, so it needs to be studied to find out whether the questions are able to measure student competence accurately.(Qomariyah, 2022) and can also measure student understanding more accurately.(Dorner et al., 2023)

One of the exam questions that needs to be analyzed for quality is an evaluation of the effectiveness of the questions as a measuring tool for Arabic language ability. Periodic item analysis of questions in order to maintain the quality of Arabic language evaluation tools. Questions have good quality grounded in the degree of challenge and differentiating capability. Evaluation of the efficacy of inquiries as a measuring tool for Arabic language ability. Periodic item analysis of questions in order to maintain the quality of Arabic language evaluation tools. Questions have good quality grounded in the level of challenge and distinguishing capability. In question analysis, there are many analysis models, one of which is the 3PL IRT Model used because it can take into account three important parameters in question analysis, namely the degree of question discrimination, the degree of challenge, and the probability of guessing the answer (guessing).(Widyana et al., 2025)

The analysis aims to determine whether the measuring instrument has been able to function as an adequate learning measurement tool or not. From the findings from the examination of each question item, it is expected that valuable information will be obtained that can be used as feedback to carry out follow-up actions on the questions that have been used in the evaluation of learning outcomes. According to the outcomes of the evaluation of a question item, several follow-ups can be carried out, whether the question item is reused as is, discarded, or improved. The evaluation is used again after the learning outcome evaluation questions can be carried out from one aspect, namely from the aspect regarding the degree of difficulty.

Before analyzing the questions, it is necessary to discuss in depth the process of compiling exam questions that are in accordance with OBE, including mapping of learning outcomes, variations in question types and difficulty levels, and fair assessments. The goal is to accurately assess attainment of student learning outcomes and facilitate a more concentrated and efficient learning experience. all components of education such as curriculum, teaching methods, and evaluations are aligned with learning outcomes. Evaluation is carried out directly (such as exams, assignments, presentations) or indirectly (such as surveys from alumni or stakeholders). Exam questions are the main component in assessing this achievement. Characteristics of Good Exam Questions include Validity:

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

Questions really measure what should be measured, Reliability: Consistency of measurement results, Objectivity: Free from personal bias, Clarity and Relevance: Easy-to-understand language, according to context and learning outcomes, and Cognitive Balance: Using Bloom's taxonomy to compile questions from low (remembering) to high (creating) levels. Bloom's taxonomy is used to compile questions based on six cognitive tiers: Remembering, Understanding, Applying, Analyzing, Evaluating, Creating. Developing exam questions based on OBE requires systematic planning and alignment with learning outcomes. This process improves the accuracy of assessment, supports effective learning, and helps universities meet accreditation standards. OBE emphasizes the achievement of measurable learning outcomes, skills, and competencies that are relevant to the world of work.(Bhandurge & Suryawanshi, 2024) Each level is accompanied by operational verbs and sample questions to make it easier for teachers to formulate appropriate questions. And in this article, we will discuss the evaluation of how challenging the final exam questions are for the Arabic Language Subject at the 'aliyah level.

In an exam, there are usually multiple choice questions and essay questions. And this time it will be discussed related to the level of difficulty of multiple choice questions. Multiple choice assessments are frequently utilized in the field of education, even at the university level stage, as an effective way to assess student comprehension. In the realm of Arabic language education, it is thought that multiple choice assessments can evaluate students' understanding and proficiency in the language. Nonetheless, assessing the practicality and dependability of these assessments is crucial to guarantee that the outcomes truly represent students' skills. This research intends to assess the practicality and dependability of multiple-choice Arabic tests in higher education. The assessment involved examining content validity, construct validity, reliability, and the correlation with students' academic performance. Assessing the practicality and dependability of Arabic multiple choice assessments. Multiple choice exams are frequently utilized in higher education as they effectively gauge student comprehension. In Arabic language learning, this test is believed in order to gauge pupils' understanding and language mastery. One of the key components of assessing the efficacy of learning for educational objectives is evaluation. Effective assessment not just offers a justification or depiction of pupils' grasp of the subject matter being instructed, but additionally assists educators in gauging and measuring the extent of student achievement in engaging with the academic program that has been implemented in college. In this instance, the The Arabic language is chosen using an appropriate and accurate methodology for assessing student competencies. One of the commonly used evaluation methods is a multiple-choice exam. This examination is often used both in schools and universities due to its capacity to deliver output data that is amenable to quantitative evaluation. As an evaluation tool for learning materials and student comprehension, multiple-choice exams are highly effective. It is crucial to make sure the test used is able to carry out an evaluation that describes the accuracy of pupils' comprehension of studying Arabic, which serves as the foundation for decisions pertaining to curriculum development and learning. Thus, the purpose of this study is to shed light on how to assess the multiple-choice exam approach in Arabic language learning.

Earlier studies by Ana Ratwa Wulan examined the fundamental nature of evaluation, assessment, testing, and measurement. The sought-after research findings in assessing for ongoing enhancement of student lessons and learning objectives are stated

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

using factual information gathered through thorough and extensive evaluations, ensuring that precise and reliable analysis of the data is possible. Consequently, there are numerous terms associated with assessments, specifically testing and measurement, which are frequently utilized by educators and instructors.

Furthermore, a study by Within the constrained evaluation domain at MTs Al-Musyawarah Lembang, Indah Rahmi et al. investigated the evaluation of Arabic language test quality with an emphasis on Higher Order Thinking Skills (HOTS). The present study employed descriptive research with a sample size of thirty people to evaluate the quality of the Arabic-language Final Semester Exam (UAS) questions. The study's findings show that the reliability and validity of the test are highly significant however, 25 questions do not comply with the standards for multiple-choice question formatting. Regarding the difficulty level achieved, there is no appropriateness, moderate discrimination ability, and adequate effectiveness of distractors.

Moreover, a study by Dina Indriana details the assessment of genuine learning and evaluation in Arabic language education. Assessment is performed with the CIPP model to assess how well learning outcomes are achieved using the approach of two key concepts in education that relate to fostering in-depth comprehension and students' capacity for critical thought. This tactic is used in both learning and evaluation. serving as a standard for the execution of learning approaches utilized by both teachers and students. The three prior studies deliver the most recent scientific insights concerning the examination of test evaluations in the suggested scientific and genuine methods to assess how accurately and precisely the test evaluation results are implemented. In this instance, The researcher used a multiple-choice technique to evaluate college students' learning of Arabic. Through performing a feasibility and reliability assessment on various options, this method presents an overview and suggestions that can effectively enhance students' grasp of Arabic language learning concepts. This study aims to discover relevant and precise feasibility and reliability findings to enhance the quality of learning Arabic.(Zakiyah et al., 2024)

This study aims to show that the BDDQ-AS's Arabic version is a reliable, legitimate self-assessment tool and is ready to use. Acceptable difficulty parameters in all items.(Abdelhamid et al., 2023) This study aims to evaluate the final semester exam questions for Aliyah class XI in order to obtain quality questions, namely questions that present a degree of challenge according to the projections at the time of question preparation and initial design. Degree of challenge (P-value): The ratio of examinees who answer questions correctly. The ideal P-value generally ranges from 0.3 to 0.7.(Law et al., 2025)

This research suitable with research conducted by Nani Fitriani written in an article entitled Analysis of Difficulty Level, Discriminating Strength, and Efficiency of Distractors for regnancy and Newborn Crisis Awareness Training Queries. Several discussions in her article related to the Degree of Challenge, namely the Degree of Difficulty of the question represents the chance to respond accurately at a particular skill level, typically indicated through an index. The difficulty index is represented as a ratio ranging from 0.00 to 1.00. The lower the number on the difficulty index, the harder the question is. Test questions can be stated as well questions if the questions are not overly challenging and not overly simple. In other terms, the degree of challenge of the questions is moderate or adequate, namely those with a difficulty index between 0.31-0.70. Sudijono showed several follow-ups which can be done upon evaluating The degree of

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

difficulty of the questions, as follows: a. Questions categorized as having a moderate level of difficulty should be stored in a question bank for future reuse. b. For questions classified as difficult, there are three potential follow-up actions: 1) The question is eliminated and will not appear again in future learning outcome assessments. 2) The inquiry is revisited to identify the elements that lead to students struggling to respond to it. Enhancements can be achieved by altering the sentence to avoid misinterpretation or substituting the figures/nominals in the calculation queries. Once the enhancements are completed, the question is reusable and can be kept in the question bank. 3) The question is kept for future use in assessments that are highly rigorous in nature, indicating that most test participants will not pass the selection test. c. For questions that are categorized as easy, there are three possible follow-up actions, namely: 1) The question is discarded and will not be issued again in the learning outcome test in the future. 2) The question is re-evaluated to determine the elements that lead nearly all students taking the exam to respond accurately. The options provided with the question might be too simple for participants to figure out. Enhancements can be achieved by refining the answer choices or increasing the complexity of the question sentence. Once the enhancements are complete, the questions may be added to the question bank and utilized in the forthcoming learning outcome assessment. 3) The questions are preserved and employed in a relaxed assessment, indicating that a majority of test takers will be considered to have succeeded in the selection. Under this circumstance, the examination is merely a formality. The findings of this research suggest that the post-test questions used have a difficulty level of "easy" of 75.55%, a "moderate" category of 8.89%, and a "difficult" category of 15.56%. The post-test questions used have a discriminatory power of mostly "bad" which is 88.89%, a "sufficient" category of 6.67%, a "good" category of 4.44% and no questions have a discriminatory power of "very good".(Fitriani, 2021)

Research conducted by Luluk Puji Rahayu and Desi Sukenti in their article entitled Quality of Indonesian Language Questions for Class XI SMAN 2 Bangko Pusako: Item Analysis. Item analysis is necessary to evaluate the quality of questions that have been developed. According to Andini & Mukhlis (2023) that to enhance the quality of questions, question makers must conduct item analysis. The goal is to enhance the quality of questions and collect data on student understanding. Supported by the opinion of Masulili et al. (2022), namely finding out about the quality of each item either through review or empirical analysis, is the main purpose of item evaluation. The results can be used to evaluate the caliber of exam questions and the caliber of the student educational process. The study's conclusions demonstrated that, based on the degree of challenge of the questions, there were 17 (57%) easy category questions, 9 (30%) medium category questions, and 4 (13%) difficult category questions.(Rahayu, 2024). Research conducted by Oky Wulandari, M. Muhtarom, and S. Sumarno in their article entitled Analysis of Mathematical Knowledge Questions for Grade V Elementary School using the Rasch framework. The study's findings revealed that eight questions were valid and two questions were invalid. The reliability obtained was categorized as quite good at 0.65. At the difficulty level of the questions, there was a representation of questions that were classified as easy, medium, and difficult.(*Jurnal Pengembangan Dan Penelitian Pendidikan*, 2025)

The questions were categorized into five levels of difficulty, from very difficult to very easy. Of the total questions, 13 (9.3%) were easy, 83 (59.3%) were moderate, and 44 (31.5%) were difficult. Questions that were too easy or too difficult were suggested

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

for re-evaluation or improvement.(Agustin, 2024) From the literature above, what distinguishes this research from previous research is the variable, namely the Arabic language subject and the place of research, namely MA Islamiyah Senori.

METHOD

This study employs a descriptive qualitative method, which aims to systematically describe, explain, and interpret data in accordance with the phenomenon being studied (Moser & Korstjens, 2018). Descriptive qualitative methods were used to describe the results of the analysis of the difficulty level of each question based on numerical data obtained through calculations using SPSS and Excel (Alvarez & Boutin, 2024). Thus, even though the initial data consisted of numerical calculations of the difficulty index, the analysis and conclusion-drawing processes were carried out qualitatively through an in-depth description of the meaning of these numbers.

The object of this study is the Semester Exam questions at MA Islamiyah Senori Class XI (20 multiple choice questions) with a total of 20 students or samples. This final exam is intended to measure learning outcomes in all aspects of Arabic language learning. Each question item is analyzed based on the answers from students, and a quantitative assessment is carried out for each question item to determine the difficulty index (P). Grammar and material substance are not included in the parameters evaluated. From the results of the quantitative evaluation, the level of difficulty is then categorized for each question item. The first step in calculating the difficulty index is to correct the participant's answer sheet. A score of 1 is given for accurate responses, whereas incorrect replies receive a score of zero. The difficulty index number (P) is calculated using equation (1) while the level of difficulty is determined using the criteria presented in the discussion.

RESULTS AND DISCUSSION

Multiple-choice questions require a choice of answers from several possible answers provided. Multiple-choice tests can be used to measure more complex learning outcomes and relate to aspects of memory, understanding, application, analysis, synthesis, and evaluation.(I. R. N. Fauziyah et al., 2020) After the researcher collected data in the form of MA Islamiyah Senori class XI Final Semester Exam questions, Answer keys, grids, question instruments, and student answers. The data was processed and analyzed, namely, by calculating the level of difficulty per item or question. In the calculation, the researcher used Excel and SPSS applications.

After being seen and examined of the outcomes of the calculation of the difficulty level of the questions totaling 20 questions with the number of students 20 students between using the SPSS application and Excel are the same. So the results of the analysis are declared strong and correct because for the initial calculation the researcher used Excel then used SPSS to strengthen it.

Exposure

The following is a presentation of the outcomes of the calculation of the difficulty level of the final exam queries for Arabic language for course XII MA Islamiyah per item or question:

Query number 1 = 0.80	Query number 11 = 0.15
Query number 2 = 0.70	Query number 12 = 0.95
Query number 3 = 0.20	Query number 13 = 0.95

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

Query number 4 = 0.60
Query number 5 = 0.85
Query number 6 = 0.70
Query number 7 = 0.70
Query number 8 = 0.45
Query number 9 = 0.65
Query number 10 = 0.65
Query number 14 = 0.45
Query number 15 = 0.95
Query number 16 = 0.80
Query number 17 = 0.95
Query number 18 = 0.85
Query number 19 = 0.70
Query number 20 = 0.40

Based on the results obtained, it can be analyzed with the formula or benchmarks of experts, including: The level of difficulty is considered effective, if a question has a moderate degree of challenge. If it is overly simple or overly challenging, it needs subject to modification. The components of the learning outcome assessment can be categorized as effective items, provided they are neither excessively difficult nor overly simple. In other terms, the level of difficulty of the item is average or adequate. Thus, the question item cannot be considered a quality item if the item cannot be answered by all testees (students) because it is too difficult, or conversely, the question item can be answered easily by all testees (students) because it is too easy. (Handriawan & Nurman, 2021). According to Djiwandono, the level of difficulty in calculations is often given the sign p. The method for calculating the level of difficulty can be obtained through simple calculations, namely with the formula.

P = (JJB:JPT) × 100%

It is known:
P = Test Item Difficulty Level
JJB = Number of Correct Answers
JPT = Number of Test Participants

According to Oller, a test item is considered feasible if its difficulty level index ranges from 0.15 to 0.85. This means that if an item index is below 0.15 or 0.10, it is considered too difficult. Conversely, if the item difficulty index is more than 0.85 or 0.90, it is considered too easy. According to Witheringthon In the book titled Psychological Education, it was mentioned that the adequacy of the learning outcome test items' difficulty can be assessed by examining the numerical value that represents the item's difficulty level. The figure that indicates the difficulty level of an item is frequently referred to as the difficulty index (item difficulty index number) and is typically represented by the letter P, representing Proportion, in the field of learning outcome assessment.

The difficulty index for the item varies from 0.00 to 1.00. The lowest difficulty is 0.00 and the highest is 1.00. If It seems like your input got cut off to 0.00 then the question item is too difficult and if the index number is nearer to 0.01 then the question item is too easy.

The formula for obtaining the index number is:

I =B:N

Information:
I = Item difficulty index
B= The number of students who answered the item correctly
N= Number of students who took the test

The criteria for the level of difficulty used are:
Items with P 0.00 to 0.30 are classified as difficult
Items with P 0.31 to 0.70 are classified as moderate

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

Items with P 0.71 to 1.00 are classified as easy

Example: Question number 1 was answered correctly by 2 test participants out of a total of 30 participants. So the way to calculate P (item difficulty index) is

P = 2:30
= 0,067

Furthermore, it can be concluded by looking at the difficulty level criteria, the number 0.067 is in the range of 0.00 - 0.30. So question number 1 is classified as difficult.

In this calculation, if the difficulty level of an item is known, then if an item is too difficult it must be revised or not used, likewise if an item is too easy it must be revised or not used. Because a suitable test is ideally not overly simple or complex.

Meanwhile, an alternative method applied to determine the difficulty level of descriptive questions is the same as multiple choice questions, namely:

Tk = $\frac{SA+SB}{IA+IB}$×100%

Tk: Difficulty level of the questions
SA: Top group score total
SB: Ideal score sum of test group
IA: Ideal score sum of test group
IB: Ideal score for lower group

Once the difficulty index is obtained, conclusions can be drawn by looking at the criteria in the table below:

Table 3. Index and Criteria Extremely Difficult

Difficulty Level Index	Criteria
0 to 15%	Extremely Challenging
16% to 30%	Difficult
31% to 70%	Moderate
71% to 85%	Simple
86% to 100%	Very Simple

Based on the benchmark using decimal numbers, namely

The criteria for the level of difficulty used are:

P values between 0.00 and 0.30 are categorized as challenging.

P values between 0.31 to 0.70 are categorized as moderate.

Items with P 0.71 to 1.00 are classified as easy

Can be concluded that The first question's degree of difficulty is 0.80 because the p value is between 0.71 and 1.00, so it is classified as easy. The second question's degree of difficulty is 0.70 because the p value is between 0.31 and 0.70, so it is classified as moderate. The level of difficulty of question number 3 is 0.20 because the p value is between 0.00 and 0.30, so it is classified as difficult. The level of difficulty of question number 4 is 0.60 because the p value is between 0.31 and 0.70, so it is classified as moderate. How challenging question number five is 0.85 because the p value is between 0.71 and 1.00, so it is classified as easy. The difficulty level of question number six is 0.70 because the value is between 0.31 and 0.70, so it is classified as moderate. The difficulty level of question number 8 is 0.45 because the p value is between 0.31 and 0.70, so it is classified as moderate. The difficulty level of question number 9 is 0.65 because the p value is between 0.31 and 0.70, so it is classified as moderate. The difficulty level of question number 10 is 0.65 because the p value is between 0.31 and 0.70, so it is classified as moderate. The difficulty level of question number 11 is 0.15 because the p

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

value is between 0.00 and 0.30, so it is classified as difficult. The difficulty level of question number 12 is 0.95 because the p value is between 0.71 and 1.00, so it is classified as easy. The difficulty level of question number 13 is 0.95 because the P value is between 0.71 and 1.00, so it is classified as easy. The difficulty level of question number 14 is 0.45 because the P value is between 0.31 and 0.70, so it is classified as moderate. The difficulty level of question number 15 is 0.95 because the P value is between 0.71 and 1.00, so it is classified as easy. The difficulty level of question number 16 is 0.80 because the P value is between 0.71 and 1.00, so it is classified as easy. The difficulty level of question number 17 is 0.95 because the P value is between 0.71 and 1.00, so it is classified as easy. The difficulty level of question number 18 is 0.85 because the P value is between 0.71 and 1.00, so it is classified as easy. The difficulty level of question number 19 is 0.70 because the P value is between 0.31 and 0.70, so it is classified as moderate. The difficulty level of question number 20 is 0.40 because the P value is between 0.31 and 0.70, so it is classified as moderate.

Thus, it can be grouped according to the findings of the analysis, namely there are 9 questions that are classified as easy, there are 9 questions that are classified as moderate, and there are 2 questions that are classified as difficult. this aligns with the outcomes of a study presented in an article titled Analysis of the Level of Difficulty of Question Items which involves evaluating questions regarding their difficulty to classify them into easy, moderate, and difficult categories. The difficulty level of a question item is obtained from related to students' ability to respond to it, not seen from the teacher's perspective in conducting analysis when compiling the question. The degree of complexity of the learning outcome evaluation query item can be seen from the size of the number that symbolizes the degree of difficulty of the question item, which is stated in the term difficulty index number, which is generally symbolized represented by the letter P, which is an can be assessed through a proportion. The difficulty index number of the question item this range moves between 0.00 and 1.00. If a a question has a difficulty index of 0.00 (P = 0.00), it means that the question is included in the category of questions that are too difficult, because none of the students allows for correct answers. Conversely, if a question has a difficulty index of 1.00 (P = 1.00), it means that the question is included in the category of questions that are too easy, because all training participants can answer the question correctly. In general, a question for evaluating learning outcomes is declared indicating that the question should not be excessively challenging or too simple for students. Consequently, questions that all students are unable to answer correctly (due to their difficulty) may be classified as bad questions. Likewise, questions that every student can answer accurately (since they are overly simple) can also be declared as bad questions. For both types of categories, improvements need to be made if they are to be used again as questions for the next exam.

The assumption used to obtain effective question quality to measure good learning outcomes is the balance of the level of difficulty of the questions. The intended balance is the comparison between the items of questions that are included in the categories of easy, medium, and difficult. The basis for determining the proportion of the number of questions in the simple, medium, and challenging sections is the purpose of the learning or test being carried out. For learning or tests that require high participant abilities, the proportion of the number of inquiries in the challenging section must be more than for learning or tests that do not require high learning outcomes. The proportion of the comparison does not have a definite value, but depends on the design and purpose of the

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

learning or test being held. The proportion is usually determined based on an agreement made when determining the design of a learning or test. After determining the proportion and level of difficulty carried out by the teachers, the questions are then tested and analyzed to determine whether the determination is in accordance with the initial design or not.

This means that in assessing student learning outcomes, most students' learning outcomes will be concentrated around the average value, and only a small number of students will have very high or very low scores. Assessments in which a participant's test results are compared with the test results achieved by other participants are called norm-referenced assessments or group-referenced assessments. If a situation occurs where the test results of the learning outcomes achieved by the training participants form an asymmetrical curve, either sloping to the left or to the right, then the examiner needs to analyze the test items that have been used as a measure of the success of the training participants.

The analysis aims to determine whether the measuring tool has been able to function as an adequate learning measurement tool or not. From the analysis results of each test item, it is hoped that valuable information will be obtained that can be used as feedback in order to follow up on the test items utilized in the evaluation of learning outcomes. Based on the outcomes of the evaluation of a test item, several follow-ups can be carried out, whether the test item is reused as is, discarded, or improved. Evaluation is used again after the learning outcome evaluation questions can be done from two aspects, namely from the difficulty level aspect and the ability to differentiate aspect. This is in line with previous research related to the analysis of final semester exam questions for Arabic for ninth-grade students at MTs Humairoh HNN Kampar for the 2024/2025 academic year. The Arabic exam consisted of 20 multiple-choice questions. The results showed that the quality of the questions was still poor because they did not meet standard standards, necessitating improvements. The validity test found 3 valid items and 17 invalid items. The reliability of the questions was categorized as low. In terms of difficulty, 15 items were classified as difficult, 4 were classified as moderate, and 1 was classified as easy. Meanwhile, in terms of discriminatory power, 2 items were categorized as good, 2 were categorized as sufficient, and 13 were categorized as weak.(Rizkiyah et al., 2025)

In the exam, it covers all the materials and skills in learning Arabic. One of them is translating. The process of changing a text from one language to another is called translation. while preserving the original meaning, intent, and tone. It is a complex and nuanced skill that requires a deep understanding of the source and target cultures, as well as linguistic proficiency.(Abu Faraj, 2024) Notable distinctions between low-level and high-level students. Low-level language learners showed a considerably more percentage of erroneous words and a reduced proportion of accurate words compared to their high-level counterparts. Further additional examination showed that beginner learners often had difficulty in identifying phonemic elements, which they often overlooked or misrepresented.(Almaiman, 2024) limited vocabulary and short implementation time. student competence in the field of Arabic translation is one of the causes.(Baihaqi et al., 2025) A possible solution is to hold online classes as additional classes. Some students feel that online classes are more effective and enjoyable than face-to-face classes, especially in terms of vocabulary mastery and self-confidence. Online language learning can be effective if it is well designed and supported by teacher training, adequate access

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

to technology, and innovative learning approaches. Virtual classes help students learn in a safe and comfortable environment, but still require supporting strategies such as active interaction, initial training for students, and continuous evaluation of learning materials and methods.(Kutubkhanah Alsaied, 2022) The proposed two-path theory-based learning adaptability model, along with the developed measurement scale and test items, offers a powerful tool for assessing and understanding students' learning adaptability. We developed a measurement scale and assessment items to evaluate students' learning flexibility more precisely.(Shi, 2025) Test item development requires calibration to determine its quality, ensuring that only high-quality items are used. High-quality questions will provide accurate information about test participants, while items of unknown quality have the potential to produce information that is inconsistent with the actual situation.(Danni et al., 2021) By evaluating the quality of test items, teachers are expected to be able to identify areas that require improvement in test construction, thereby increasing the overall effectiveness of the learning process.(Muslimah & Widiyanti, 2023)

CONCLUSION

Evaluation is an essential component of every learning process because it determines the extent to which students have achieved the intended competencies. One of the key elements in evaluation is the quality of the assessment instruments, especially test items. To obtain accurate evaluation results, each item must be analyzed particularly in terms of its difficulty level, which classifies questions into easy, moderate, or difficult. This analysis is crucial to ensure that ineffective items can be revised or removed for future assessments. In the context of Arabic as a foreign language, item difficulty analysis becomes even more important. Indonesian students are not native speakers of Arabic, and variations in vocabulary mastery, reading comprehension, and understanding of linguistic structures can greatly influence their test performance. Therefore, this study specifically aims to examine the difficulty level of the Arabic Final Semester Examination items for Grade XI students at MA Islamiyah Senori. Through this analysis, the research seeks to determine whether the test items align with learning objectives and whether they accurately measure students' abilities. The findings are expected to serve as a basis for improving the quality of Arabic language assessment instruments in subsequent evaluations.

REFERENCES

Abdelhamid, A. S., Elzayat, S., Amer, M. A., Elsherif, H. S., Lekakis, G., & Most, S. P. (2023). Arabic translation, cultural adaptation, and validation of the BDDQ-AS for rhinoplasty patients. *Journal of Otolaryngology - Head and Neck Surgery*, *52*(1), 1–7. https://doi.org/10.1186/s40463-022-00613-6

Abu Faraj, B. R. (2024). Mona Baker's Strategies Used for Translating the Arabic HAND Idioms. *Theory and Practice in Language Studies*, *14*(1), 139–145. https://doi.org/10.17507/tpls.1401.16

Agustin, B. I. (2024). *Analysis of Item Difficulty Levels and Differentiating Power of TOAFL Questions at KH . Abdul Wahab Hasbullah University Jombang*. *9*(2).

Almaiman, I. (2024). Spelling Difficulties among EFL Students: An Error Analysis Framework Using Computer Software- the Spelling Sensitivity Score (SSS).

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license (https://creativecommons.org/licenses/by-sa/4.0/)

World Journal of English Language, *14*(4), 437–446. https://doi.org/10.5430/wjel.v14n4p437

Alvarez, J. D. O., & Boutin, M. (2024). *Beyond analytics : Using computer - aided methods in educational research to extend qualitative data analysis*. *November 2023*. https://doi.org/10.1002/cae.22749

Baihaqi, M., Syarifah, A., & Arif, M. (2025). Enhacing Arabic Translation Competency In Higher Education: An Evaluation Of The Becoming A Translation Practitioner' Program. *Ijaz Arabi: Journal of Arabic Learning*, *8*(1), 493–503.

Bhandurge, P., & Suryawanshi, S. S. (2024). Question Paper Design in Line with Outcome-Based Education Policy. *Indian Journal of Pharmaceutical Education and Research*, *58*(4), s1201–s1210. https://doi.org/10.5530/ijper.58.4s.117

Danni, R., Wahyuni, A., & Tauratiya. (2021). Item Response Theory Approach : Kalibrasi Butir Soal Penilaian Akhir Semester Mata Pelajaran Bahasa Arab. *Arabi: Journal of Arabic Studies*, *6*(1), 93–104.

Dianova, F. R., & Anwar, N. (2024). Analisis Butir Uji Validitas , Reliabilitas , Tingkat Kesukaran , dan Daya Pembeda Soal Sumatif Bahasa Arab SD Islam. *Jurnal Bahasa Daerah Indonesia*, *3*, 1–13.

Dorner, M. A., Sadler, P., & Alters, B. (2023). Still a private universe? Community college students' understanding of evolution. *Evolution: Education and Outreach*, *16*(1), 1–18. https://doi.org/10.1186/s12052-022-00178-y

Fauziyah, I. R. N., Syihabudin, & Sopian, A. (2020). Analisis Kualitas Tes Bahasa Arab Berbasis Higher Order Thinking Skill (HOTS). *Lisanuna*, *10*(1), 45–54.

Fauziyah, L., Masitho, D. N., Baihaqi, M., Syafitri, N. S., & Yusuf, H. D. (2025). Idhmaju Taqyiimi al-Ashili Fi Ta ' allumi Lughotil ' Arabiyyat: al-Mandzuratu Nadzoriyyat Wal- Athari Ghoir al-Ikhtibariyyah. *Al Mahara: Jurnal Pendidikan Bahasa Arab*, *11*(2), 330–353. https://doi.org/10.14421/almahara.2025.

Fitriani, N. (2021). Analisis Tingkat Kesukaran, Daya Pembeda, Dan Efektivitas Pengecoh Soal Pelatihan Kewaspadaan Kegawatdaruratan Maternal Dan Neonatal. *Paedagoria : Jurnal Kajian, Penelitian Dan Pengembangan Kependidikan*, *12*(2), 199. https://doi.org/10.31764/paedagoria.v12i2.4956

Handriawan, D., & Nurman, M. (2021). Evaluasi Pembelajaran Bahasa Arab. In *Sanabil Publishing*. Sanabil.

Jundi, A., Baihaqi, M., & Zaenuri, M. (2024). Identification And Correction of Pseudowords in Ilman Wa Ruhan Textbooks to Reduce Meaning Errors. *Ijaz Arabi:Journal of Arabic Learning*, *8*(1), 0–11.

Jurnal Pengembangan dan Penelitian Pendidikan. (2025). *07*(1), 293–302.

Khan, H. F., Qayyum, S., Beenish, H., Khan, R. A., Iltaf, S., & Faysal, L. R. (2025). Determining the alignment of assessment items with curriculum goals through document analysis by addressing identified item flaws. *BMC Medical Education*, *25*(1). https://doi.org/10.1186/s12909-025-06736-4

Kutubkhanah Alsaied, H. I. K. (2022). Impact of Distance Learning on the English Language Learning Process. *Journal of Curriculum and Teaching*, *11*(7), 37–47. https://doi.org/10.5430/JCT.V11N7P37

Law, A. K. K., So, J., Lui, C. T., Choi, Y. F., Cheung, K. H., Kei-ching Hung, K., &

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

Graham, C. A. (2025). AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Medical Education*, *25*(1). https://doi.org/10.1186/s12909-025-06796-6

Moser, A., & Korstjens, I. (2018). Series : Practical guidance to qualitative research . Part 3 : Sampling , data collection and analysis. *European Journal of General Practice*, *0*(0), 9–18. https://doi.org/10.1080/13814788.2017.1375091

Muslimah, M., & Widiyanti, A. (2023). Analisis Daya Beda Tes Hasil Belajar Bahasa Arab Siswa SMA Mamba ' ul Hikmah Paron Ngawi. *Al-Mu'arrib: Jurnal Pendidikan Bahasa Arab*, *3*(2), 67–77. https://doi.org/10.32923/al-muarrib.v3i2.3594

Qomariyah, L. (2022). Analisis Tingkat Kesukaran dan Daya Pembeda Butir Soal TOAFL Universitas Hasyim Asy'ari Tebuireng Jombang. *Lisanan Arabiya: Jurnal Pendidikan Bahasa Arab*, *6*(1), 1–18. https://doi.org/10.32699/liar.v6i1.2549

Rahayu, L. P. (2024). *Kualitas Soal Bahasa Indonesia Kelas XI SMAN 2 Bangko Pusako : Analisis Butir Soal*. *10*(4), 3755–3762.

Rizkiyah, A. M., Hikmah, & Masrun. (2025). Analisis Kualitas Tes dan Butir Soal Ujian Madrasah Mata Pelajaran Bahasa Arab Kelas IX. *Psikosospen: Jurnal Psikososial Dan Pendidikan*, *1*(2), 533–543.

Shi, X. (2025). Study on learning adaptability of STEM students in the new college entrance examination context. *Discover Education*, *4*(1). https://doi.org/10.1007/s44217-025-00425-6

Widyana, R., Kadir, P. M., & Masruria, W. W. (2025). *Three-Parameter Item Response Theory Analysis of the Multiple-Choice Items in PIRLS 2016*. *15*(2).

Zakiyah, Sulfiatin, Baihaqi, M., Faiza, N. M., Alghani, M. I. R., & Fatihuddin. (2024). Evaluation of the Feasibility and Reliability of Arabic Multiple Choice Test in Higher Education. *Lisanudhad: Jurnal Bahasa, Pemeblajaran Dan Sastra Arab*, *11*(2), 164–192.

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)