

Analysis of the Difficulty Level and Discriminating Power of Arabic Language Test Items at MTS Salafiyah Merakurak Tuban

Siti Nur Salma Salsabela¹, M. Baihaqi², Mohammad Nu'man³

^{1,2,3}Unicersitas Islam Negeri Sunan Ampel Surabaya, Indonesia

Email: salmapba23@gmail.com

Abstract

The quality of test items determines how accurately an assessment represents students' ability, yet item analysis of Arabic language tests in madrasahs remains uncommon, although teachers need its results to improve their tests. This study aimed to describe the item difficulty and discrimination index of 50 items from a ninth-grade end-of-semester Arabic language examination at MTs Salafiyah Merakurak Tuban. A descriptive quantitative approach was applied to the dichotomously scored responses of 20 students. Item difficulty was calculated as the proportion of students answering correctly, and the discrimination index as the difference between the mean scores of the upper and lower groups. The results showed that 23 items (46%) were easy, 20 items (40%) moderate, and 7 items (14%) difficult. The mean discrimination index was 0.104; 6 items (12%) had negative values, 33 items (66%) were poor, and only 11 items (22%) were good or very good. Although the difficulty distribution was fairly varied, the items' power to distinguish between students remained weak, so most items need to be reviewed, revised, or replaced before being stored in an item bank.

Keywords: *Item Analysis, Item Difficulty, Discrimination Index, Arabic Language Assessment, Educational Evaluation.*

A. INTRODUCTION

Evaluation is one of the important parts of a learning process. Evaluation is the process of collecting and processing information carried out systematically to assess the achievement of a learning objective, determine the effectiveness of learning, and serve as a basis for accountable decisions in education (Khaled & Salameh, n.d.). Through the analysis of learning evaluation, teachers gain knowledge about effective strategies to implement, understand students' learning abilities, and design learning strategies to carry out appropriate follow-up actions in subsequent learning activities (Zuhri & Syihabuddin, 2024). Without evaluation, teachers do not know the extent to which students have mastered the competencies that have been taught, resulting in less structured learning (Baihaqi et al., 2025).

In learning evaluation, the most commonly used instrument is the learning achievement test, which is designed to measure students' competencies in the material that has been taught (Magdalena et al., n.d.). Learning achievement tests in the form of multiple-choice questions are widely chosen because they are practical and can cover a broad range of cognitive domains (Lailatul Qomariyah, Isnihatun Niswah, 2021). Using tests does not guarantee valid and reliable data. The quality of a test is determined by the quality of its items; test items that do not meet psychometric standards can produce inaccurate data and mislead teachers' decisions

(Ahmad Wildan Sahuri Ramdani Muammar Qadavi Umasugi Ubaid Ridlo Raswan, 2025). Therefore, each test item must be analyzed to ensure that the questions are valid, accurate, and able to measure students' competencies effectively, thereby revealing students' actual abilities (Leonardus Kaka, Vidriana Oktoviana Bano, 2024). Item analysis is essentially the process of examining each test item to identify good and poor items (Ramadhan, 2024) by correlating the score of each item with the overall test score (Fikriyah, 2021).

Learning evaluation is important in Arabic language learning because evaluation determines the quality of learning and students' success in mastering the language (Nafisah, 2024). Item analysis of Arabic language tests has been conducted at various educational levels and across different types of tests, ranging from Islamic elementary schools, junior high schools and MTs, Islamic senior high schools, to intensive programs at higher education institutions (Ambiya, 2022).

Two commonly used indicators in item analysis are item difficulty and discrimination power. Item difficulty indicates how easy or difficult an item is for a group of students, as measured by the proportion of students who answer correctly (Fauziah, 2020). A good instrument should contain a balanced distribution of easy, moderate, and difficult questions (Damianus Nd. Marambaawang, Vidriana Oktoviana Bano, 2023). Questions that are too easy tend to be less effective in measuring the abilities of high-performing students, while questions that are too difficult may reduce motivation and result in disproportionate assessments (Zahra et al., 2026). Discrimination power indicates an item's ability to distinguish between high-ability and low-ability students (Hera Apriliana Saputri, Zulhijrah, Nabila Joti Larasati, 2023). High-quality items tend to be answered correctly by students who have mastered the material and incorrectly by students who have not mastered it. Conversely, low or negative discrimination power indicates that an item does not function as intended (Rostianingsih, M.Baihaqi, 2025). Both indicators are interrelated and are generally analyzed together to determine item characteristics based on students' responses (Shakurnia et al., 2022).

Previous research on Arabic language test items has revealed varied patterns. (Azdiva Miftahur Rizkiyah & Masrun, 2025) analyzed 20 items from a ninth-grade madrasah examination at an MTs involving 14 students and found 15 difficult items, 4 moderate items, and 1 easy item, with 13 items showing weak discrimination power, 2 showing fair discrimination power, and 2 showing good discrimination power. In contrast, (Yumna Hanifa Abdillah, 2024) reported that, in the eleventh-grade test at an Islamic senior high school, 50% of the items were easy, 30% were very easy, 20% were moderate, and none were difficult, with 40% of the items demonstrating very good discrimination power and 20% demonstrating good discrimination power. (Rostianingsih, M.Baihaqi, 2025) examined Arabic language final semester examination questions and reported a predominance of moderate difficulty levels. These differences indicate that the distribution of item difficulty and discrimination power depends heavily on the educational level, type of test, and test developer. Therefore, findings from one institution cannot simply be applied to

another institution.

Based on a review of previous studies, no research has been identified that specifically analyzes Arabic language test items at MTs Salafiyah Merakurak Tuban. This study was conducted to fill this research gap through three aspects. First, item difficulty and discrimination power are analyzed together so that the relationship between the two indicators can be examined, rather than merely presenting the results of each indicator separately. Second, the research object is the ninth-grade Arabic language Final Semester Examination (UAS) for the 2023/2024 academic year, which reflects the characteristics of the madrasah, particularly in the implementation of learning materials based on Islamic values. Third, the data were obtained directly from the actual examination, rather than from simulated test questions, so that the results reflect real conditions in the field. The purpose of this study is to describe (1) the item difficulty level and (2) the discrimination power of the Arabic language Final Semester Examination items at MTs Salafiyah Merakurak Tuban.

B. METHOD

This study employed a quantitative approach with a descriptive design because it aimed to describe the empirical characteristics of each test item without examining relationships or effects between variables. The study was conducted at MTs Salafiyah Merakurak Tuban. The object of the study consisted of 50 items designed for the ninth-grade Arabic Language Final Semester Examination (UAS) for the 2023/2024 academic year, and the data were obtained from the answers of 20 students who took the examination. The UAS consisted of multiple-choice questions. The data were obtained from the documentation of the examination results. Correct answers were assigned a score of 1, while incorrect answers were assigned a score of 0. The scores were then tabulated in a matrix of students and test items.

The difficulty level of each item can be calculated by comparing the number of students who answered correctly (B) with the total number of students (JS), using the formula $P = B/JS$ (Tanjung et al., 2024). To determine discrimination power, students were ranked based on their total scores and then divided into the upper group (10 students) and the lower group (10 students). The proportions of students who answered correctly in the two groups were then compared (Riswanda Himawan, 2022). The discrimination index (D) is the difference between the average scores of the upper group (PA) and the lower group (PB), expressed as $D = PA - PB$, with a range from -1.00 to +1.00 (Faradisi & , Mahmudah , M. Baihaqi , Ajeng Nadlifah, 2025).

The interpretation of the item difficulty index follows the criteria of (A.M.S. et al., 2025), which are presented in Tables 1 and 2. This source examines the quality of multiple-choice questions in a medical faculty; however, the indicators used are general psychometric measures, namely the proportion of correct answers and the differences between groups. Therefore, these indicators can be applied to multiple-choice questions in other subjects, including Arabic. A difficulty index of 0.50–0.60 is considered ideal, while an index of 0.30–0.70 is still acceptable. The discrimination

power criteria used are more flexible than those of other criteria, with a threshold of 0.15 for good discrimination and 0.25 for very good discrimination. Therefore, readers should interpret the results while taking these criteria into consideration.

Table 1. Item Difficulty Criteria

Index (P)	Category	Meaning
$P < 0,30$	Difficult	May reduce motivation if excessive
$0,30 \leq P \leq 0,70$	Moderate (acceptable)	Ideal range: 0.50–0.60
$P > 0,70$	Easy	Less informative for high-ability students

Table 2. Discrimination Power Criteria

Index (D)	Category	Follow-up Action
$D < 0,00$	Negative (problematic)	Thoroughly reviewed
$0,00 \leq D < 0,15$	Poor	Discarded or revised
$0,15 \leq D \leq 0,24$	Good	Retained
$D \geq 0,25$	Very good	Retained

The scope of this study is limited to analyzing item difficulty and discrimination power. Validity, reliability, and distractor effectiveness were not analyzed; therefore, conclusions regarding item quality apply only to these two indicators.

C. RESULTS AND DISCUSSION

1. Item Difficulty Level

Item difficulty is a basic indicator for determining the quality of each test item, which indicates how easy or difficult an item is for a particular group of test-takers (Azdiva Miftahur Rizkiyah & Masrun, 2025). The lower the percentage of test-takers who answer correctly, the more difficult the item, and vice versa (Yumna Hanifa Abdillah, 2024). A variety of difficult questions should still be retained to distinguish different levels of competence; however, an excessive proportion of difficult questions may reduce students' motivation and result in disproportionate assessments (Najwa Zalfa Zuhri, Syihabuddin, 2024). Conversely, having too many easy questions may fail to provide sufficient information about students' mastery at higher levels.

Based on the criteria in Table 1, the results of the analysis of the 50 test items are presented in Table 3. The average difficulty index of all items was 0.63.

Table 3. Distribution of Item Difficulty Levels (N = 50 items; 20 students)

Category	Number of Items	Percentage
Easy	23	46%
Moderate	20	40%
Difficult	7	14%
Total	50	100%

Easy items: 1, 2, 3, 6, 7, 15, 16, 19, 20, 22, 23, 24, 25, 26, 35, 36, 38, 41, 43, 46, 47, 49, and 50. Difficult items: 14, 29, 32, 33, 34, 37, and 45. All other items are categorized as moderate, namely items 28 and 30, whose indices were calculated

after correcting the score codes.

Easy items are the most dominant category, but their proportion differs from that of moderate items by only 6 percentage points. This indicates that the questions tend to be easy to moderate, suggesting that they are not overly challenging for students (Sofiani Putri Radja, Vidriana Oktoviana Bano, 2023). Six items (12%) fall within the ideal range of 0.50–0.60, namely items 4, 11, 18, 21, 42, and 48. The tendency of the instrument is particularly evident at the upper end of the scale: 17 out of 50 items have an index of 0.90 or higher, and items 19 and 46 were answered correctly by all students (index 1.00). At the lower end, items 14 and 29 were not answered correctly by any student (index 0.00), while items 33 and 37 were answered correctly by only one student (index 0.05).

Extreme values need to be interpreted carefully. An item is considered very difficult when no students answer it correctly; however, this may also result from an incorrect answer key, material that has not been taught, or misleading wording. Conversely, having many easy items does not necessarily indicate a weakness. If the UAS is intended to assess the mastery of basic competencies, a high percentage of correct answers may reflect the success of the learning process.

Compared with previous studies, this distribution lies between two contrasting patterns. The dominance of easy items was also reported by Abdillah dan Hikmah (2024), whose previous study found no difficult items. However, the proportion of easy items in the present study is lower, while the proportion of moderate items is higher (Yumna Hanifa Abdillah, 2024). In contrast, (Yumna Hanifa Abdillah, 2024) found that 15 out of 20 items in a ninth-grade MTs examination were difficult, placing the instrument at MTs Salafiyah Merakurak at the opposite end of the spectrum. The findings of Rostianingsih et al. (2025), which showed a predominance of moderate items, are closer to the ideal distribution. Thus, this instrument has a reasonably varied distribution, but it is not yet fully balanced.

For teachers and test developers, this study provides three main findings. Items with moderate difficulty levels can be prioritized for inclusion in the question bank after their discrimination power has been evaluated. Easy items, particularly those with indices of 0.90 or higher, should be reviewed to determine whether they are appropriate or whether their cognitive demands could be increased. Difficult items require a review of the answer key and the clarity of their wording before revision. In addition, increasing the number of moderately difficult items through the systematic development of a test blueprint may help achieve a more balanced distribution of difficulty levels.

2. Item Discrimination Power

Discrimination power indicates an item's ability to distinguish between students with high levels of mastery and those with low levels of mastery based on their examination scores. Effective items are more frequently answered correctly by the upper group than by the lower group, making them effective diagnostic instruments for identifying students' levels of mastery (Muslimah & Widiyanti,

2023). A negative value indicates the opposite pattern: lower-performing students answer correctly more frequently than higher-performing students, and such items are classified as problematic (Faradisi & , Mahmudah , M. Baihaqi , Ajeng Nadlifah, 2025). Discrimination power results are also used to determine whether an item can be used, needs to be revised, or should be rejected (Sofiani Putri Radja, Vidriana Oktoviana Bano, 2023).

In this study, the 20 students were divided into upper and lower groups, each consisting of 10 students. With such a small group size, a difference of one student changes the index by 0.10, while moving one student from one group to the other can change the index by 0.20. Five students with the same score (32) were positioned at the boundary between the two groups, so their placement could affect the results.

Table 4. Distribution of Item Discrimination Power (N = 50 items)

Category	Index Range	Number of Items	Percentage
Negative	$D < 0,00$	6	12%
Poor (low)	$0,00 - < 0,15$	33	66%
Good	$0,15 - 0,24$	3	6%
Very good	$\geq 0,25$	8	16%
Total		50	100%

The average discrimination index was 0.104. Negative items: 24, 28, 34, 37, 43, and 45. Good items: 21, 23, and 39. Very good items: 5, 8, 9, 10, 17, 18, 27, and 32. The average discrimination index of 0.104 is considered low; however, the average alone is insufficient to assess the instrument, so the distribution of items needs to be examined. A total of 39 items (78%) did not reach 0.15, including 6 negative items, and 11 of these items had an index of exactly 0.00. Only 11 items (22%) were categorized as good or very good. This pattern is consistent with Rizkiyah et al. (2025), who recorded 13 out of 20 items with weak discrimination power. Abdillah and Hikmah (2024), who reported that 40% of items had very good discrimination power and 20% had good discrimination power, demonstrated that low discrimination power is not inevitable in Arabic language tests, but rather a matter of item quality and test characteristics.

Part of the explanation lies in the relationship with item difficulty (Table 5). Of the 11 items categorized as good or very good, 9 were of moderate difficulty, while item 23 (easy) and item 32 (difficult) accounted for one item each. Arithmetically, items answered correctly by all students (items 19 and 46) or items that no students answered correctly (items 14 and 29) must have a discrimination index of 0.00, while other easy items leave little room for the upper group to outperform the lower group. Consequently, 20 of the 23 easy items had poor discrimination power, with an average index of only 0.07. Difficult items had an average index of -0.01 , and three of the seven difficult items had negative values (items 34, 37, and 45). The average index for moderate items reached 0.19. These findings reinforce the rationale for having a test distribution dominated by moderately difficult items.

**Table 5. Relationship Between Item Difficulty and Discrimination Power
(Number of Items)**

Item Difficulty Category	Negative	Poor	Good	Very Good	Total
Easy	2	20	1	0	23
Moderate	1	10	2	7	20
Difficult	3	3	0	1	7
Total	6	33	3	8	50

The six negative items require specific examination. Items 24, 37, and 43 have values of -0.10 , equivalent to a difference of one student, meaning that these results may occur by chance in such a small group. Items 28 and 34 (-0.20), and especially item 45 (-0.30), indicate stronger evidence of problems and should therefore be prioritized for review, starting with the accuracy of the answer key, clarity of wording, and the possibility of errors in the scoring process. The specific causes cannot be determined because distractor selection data are unavailable, even though distractor analysis is a way to explain the relationship between difficulty and discrimination power (Shakurnia et al., 2022). This study can only identify the problematic items, not explain why the lower group answered correctly more frequently.

Two points need to be addressed before the discrimination power categories are finalized. First, the remarks column in the discrimination power file labels all 50 items as “poor,” even though 11 items have indices of 0.15 or higher and are categorized as good or very good according to Table 2. This article uses categories based on numerical values. Second, discrimination power in such small groups is highly sensitive to group classification: correcting the score codes shifts two students between groups and changes some indices by as much as 0.40. If stricter criteria than those of Eleragi et al. (2025), are applied, such as thresholds of 0.30 or 0.40, even fewer items meet the criteria (8 items with indices ≥ 0.30 and 6 items ≥ 0.40).

Follow-up actions can be determined based on Table 2 and the difficulty indices, as summarized in Table 6. Moderately difficult items with very good discrimination power are suitable for forming the core of the question bank, negative items require thorough examination, and items with extreme indices require a review of their answer keys. For items with low discrimination power, teachers are advised to review the content, language, and answer choices to ensure that the questions function as intended (Zahra et al., 2026). Most poor items belong to the easy category; such items should not automatically be discarded but should instead be reviewed to determine whether they are still necessary in examinations intended to distinguish between students’ abilities (Baihaqi et al., 2024).

Table 6. Recommended Follow-Up Actions Based on the Analysis Results

Item Group	Item Numbers	Recommendation
Moderate difficulty, very good discrimination ($D \geq 0.25$)	5, 8, 9, 10, 17, 18, 27	Retain and include in the question bank

Difficult, very good discrimination	32	Retain; check the answer key because $P = 0.20$
Good discrimination (0.15–0.24)	21, 23, 39	Retain; review the wording and answer choices
Negative discrimination ($D < 0$)	24, 28, 34, 37, 43, 45	Check the answer key, wording, and scoring; prioritize items 45, 28, and 34
Extreme difficulty index of 0.00 or 1.00	14, 29 (0,00); 19, 46 (1,00)	Check the answer keys for items 14 and 29; review the usefulness of items 19 and 46
Other items with poor discrimination ($D = 0.00$ –0.10)	All items not listed above	Revise or replace according to the examination objectives

Overall, this instrument demonstrates a reasonably varied distribution of item difficulty, but its ability to discriminate between students remains weak. Consequently, the student rankings produced by this UAS are less reliable as a basis for distinguishing levels of ability, although the scores can still indicate the achievement of certain learning materials. Test improvement should prioritize items that show negative results, as well as moderately difficult items that have the potential to achieve better quality after revision. This study has several limitations: the number of participants was limited to 20 students, the indices were calculated based on only one examination, and validity, reliability, and distractors were not analyzed. Future research should involve a larger number of participants and incorporate distractor analysis so that the causes of low discrimination power can be explained rather than merely identified.

D. CONCLUSION

The difficulty levels of the 50 Arabic Language Final Semester Examination (UAS) items for Grade IX at MTs Salafiyah Merakurak Tuban were distributed as follows: 23 easy items (46%), 20 moderate items (40%), and 7 difficult items (14%). The test tended toward easy items, with 17 items having an index of 0.90 or higher, while the proportion of moderate items remained relatively substantial. Discrimination power was classified as weak, with an average index of 0.104; 6 items (12%) had negative discrimination, 33 items (66%) had poor discrimination, 3 items (6%) had good discrimination, and 8 items (16%) had very good discrimination. Most items with adequate discrimination power belonged to the moderate difficulty category. Overall, item quality in terms of difficulty and discrimination power was not yet optimal; most items need to be reviewed, revised, or replaced before being included in the question bank.

Teachers are advised to retain moderately difficult items with good discrimination power, check the answer keys for items with extreme indices and negative discrimination, and increase the number of moderate items through a carefully planned test blueprint. Test developers are advised to make item analysis a routine stage after each examination. Future research should involve a larger sample and include analyses of validity, reliability, and distractor effectiveness.

REFERENCES

1. Abdillah, Y. H., & Hikmah, K. (2024). Analisis butir soal bahasa Arab siswa kelas XI MA Islamiyah Sidoarjo berbasis software anates versi 4.0. 9. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, 9(04), 375-391.
2. Ambiya, I. Z. (2022). Analisis Soal Ujian Mata Pelajaran Bahasa Arab Kelas 12 di Madrasah Aliyyah. *EL-Ibtikat Jurnal Pendidikan Bahasa Arab*, 11(1), 70–87.
3. Baihaqi, M., Mahmudah, M., Al Mubarak, A. J., Rabbaniyah, N., & Susanto, D. (2025). Studi Validasi dan Reliabilitas Butir Soal Tes Sumatif dalam Pendidikan Bahasa Arab. *Cerdika: Jurnal Ilmiah Indonesia*, 5(4), 1530-1542.
4. Baihaqi, M., Musyafaul, N., Ilham, M., & Alghani, R. (2024). Evaluation of the Feasibility and Reliability of Arabic Multiple-Choice Test in Higher Education. *Lisanudhad: Jurnal Bahasa, Pembelajaran dan Sastra Arab*, 11(2), 164–192.
5. Eleragi, A. M., Miskeen, E., Hussein, K., Rezigalla, A. A., Adam, M. I., Al-Faifi, J. A., ... & 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), 233.
6. Faradisi, R. J., Mahmudah, M., Baihaqi, M., Nadlifah, A., & Mufidah, I. (2025). Kualitas Butir Soal dalam Evaluasi Bahasa Arab: Kajian Empiris pada PAS Program Intensif. *An-Nidzam: Jurnal Manajemen Pendidikan dan Studi Islam*, 12(1), 24-39.
7. Fauziah, I. R. N., Syihabudin, S., & Sopian, A. (2020). Analisis Kualitas Tes Bahasa Arab Berbasis Higher Order Thinking Skill (Hots). *LISANUNA: Jurnal Ilmu Bahasa Arab Dan Pembelajarannya*, 10, 45-54.
8. Fikriyah, N. (2021). Analisis Butir Soal Ulangan Tengah Semester Mata Pelajaran Bahasa Arab Kelas VII Semester Ganjil SMP Muhammadiyah 1 Yogyakarta Tahun Ajaran 2019 / 2020. *Maharaat: Jurnal Pendidikan Bahasa Arab*, 3(2), 128–140. <https://doi.org/10.18196/mht.v3i2.10501>
9. Kaka, L., Bano, V. O., & Njoeroemana, Y. (2024). Efektivitas analisis butir soal pilihan ganda menggunakan aplikasi anates di SMPN 2 Kanatang. *Jurnal Inovasi Penelitian*, 4(9), 1441-1450.
10. Magdalena, I., Anggraini, I. A., & Khoiriah, S. (n.d.). Analisis Daya Pembeda, Dan Taraf Kesukaran Pada Soal Bilangan Romawi Kelas 4 SDN Tobat 1 Balaraja. *Nusantara: Jurnal Pendidikan dan Ilmu Sosial*, 3, 151–158.
11. Marambaawang, D., Bano, V. O., & Enda, R. R. H. (2023). Analisis Kualitas Butir Soal Penilaian Akhir Semester Gasal Tahun 2021/2022 Menggunakan Iteman di SMP Negeri 1 Kambara. *Dharmas Education Journal (DE_Journal)*, 4(1), 233-243.

12. 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>
13. Nafisah, F. A. M. N. (2024). Kajian Evaluasi Pembelajaran Bahasa Arab di Perguruan Tinggi (Teori dan Praktik). *EL-IBTIKAR: Jurnal Pendidikan Bahasa Arab*, 13(1), 19–27.
14. Qomariyah, L., & Isniyatun Niswah, M. (2021). Analisis Kualitas Butir Soal Toafl Unhasy Tebuireng. In *Prosiding Seminar Nasional Sains, Teknologi, Ekonomi, Pendidikan dan Keagamaan (SAINSTEKNOPAK)* (Vol. 5, No. 1).
15. Radja, S. P., Bano, V. O., & Ina, A. T. (2023). Analisis kualitas butir soal hasil belajar peserta didik berdasarkan tingkat kesukaran, daya beda dan efektivitas pengecoh di SMAN 1 Pandawai. *Jurnal Edusavana*, 1(1), 30-41.
16. Ramadhan, S. (2024). *Analisis Butir Soal Posttest Pada Materi Lingkaran Dengan Kemampuan Pemecahan Masalah dan Kemandirian Belajar*.
17. Riswanda Himawan, B. N. (2022). Analisis butir soal latihan penilaian akhir semester ganjil mata pelajaran bahasa Indonesia kelas VIII SMPN 1 Bambanglipuro Bantul menggunakan program ITEMAN (Analysis of exercise items for odd semester end of semester Indonesian language subjects class. *KEMBARA: Jurnal Keilmuan Bahasa, Sastra, Dan Pengajarannya*, 8(1), 160–180.
18. Rizkiyah, A. M. (2025). Analisis Kualitas Tes dan Butir Soal Ujian Madrasah Mata Pelajaran Bahasa Arab Kelas IX. *Jurnal Psikososial dan Pendidikan*, 1(2), 533-543.
19. Rostianingsih, M., & Baihaqi, R. (2025). Item Analysis of Multiple-Choice Questions of Arabic Language Subjects in the Final Semester Exam. *Ethical Lingua*, 12(02), 464–474. <https://doi.org/10.30605/25409190.855>
20. Salameh, E. K. M., & Alkhateeb, N. A. M. (2025). The Effectiveness of Online Education on the Development of Critical Thinking Skills among Nursing Students in Al-Balqa Applied University: A Quasi-Experimental Study. *Journal of Educators Online*, 22(2), 2.
21. Saputri, H. A. S., & Larasati, N. J. (2023). Analisis instrumen assesmen: validitas, reliabilitas, tingkat kesukaran dan daya beda butir soal. *Didaktik: Jurnal Ilmiah PGSD STKIP Subang*, 9(5), 2986-2995.
22. Shakurnia, A., Ghafourian, M., Khodadadi, A., Ghadiri, A., Amari, A., & Shariffat, M. (2022). Evaluating Functional and Non-Functional Distractors and Their Relationship with Difficulty and Discrimination Indices in Four-Option Multiple-Choice Questions. *Education in Medicine Journal*, 14(4), 55–62. <https://doi.org/10.21315/eimj2022.14.4.5>
23. Tanjung, M. A. H. R., Fahmi, A. A., Habibah, F. R. I. F., Ihda, & Qomari, N. (2024). Analisis Butir Soal Penilaian Akhir Tahun Pelajaran Bahasa Arab Kelas VII MTs Al-Ma'arif Rakit Banjarnegara Jawa Tengah. *Mantiqu Tayr: Journal of Arabic Language*, 4(1).
24. Wahdi, S., Syuhada, M. R. A., & Ridlo, U. (2025). Pembelajaran bahasa Arab berbasis higher order thinking skills (HOTS) pada kurikulum merdeka. *Jurnal Pendidikan Bahasa Arab*, 3(2), 137-152.

25. Zahra, M. J., Salim, A., & Abdurahman, I. (2026). Analisis Butir Soal Ujian Tengah Semester Mata Kuliah al- Tadrībāt di Ma'had al -Imarat Putri Bandung. *Al Matla: Journal of Arabic Language Education*, 1(1), 25–40.
26. Zuhri, N. Z., Syihabuddin, S., & Tatang, T. (2024). Analisis Validitas, Reliabilitas, dan Tingkat Kesukaran Soal Bahasa Arab Tingkat SMP Berbasis Artificial Intelligence (AI) melalui Platform QuestionWell. *Jurnal Pendidikan dan Pembelajaran Indonesia (JPPI)*, 4(2), 693-704.