AL-LISAN: JURNAL BAHASA
Publisher: LPPM IAIN Sultan Amai Gorontalo
ISSN: 2442-8965 E-ISSN: 2442-8973
Volume 11, No. 1 February 2026
Journal Homepage: https://journal.iaingorontalo.ac.id/index.php/al

Quality Profile of Arabic Final Semester Assessment Items: A Psychometric Analysis

[1]Zuliyah Safitri (Corresponding Author)
zuliyahsafitri14@gmail.com
Arabic Language Education, Faculty of Tabiyah and Teacher Training, Sunan Ampel State Islamic University, Indonesia

[2]M. Baihaqi
baihaqi@uinsa.ac.id
Arabic Language Education, Faculty of Tabiyah and Teacher Training, Sunan Ampel State Islamic University, Indonesia

ABSTRACT

Background: The quality of assessment instruments is essential to ensure that students' learning outcomes are measured accurately. In Arabic language learning, Final Semester Assessments (PAS) must be supported by sound psychometric qualities to function as valid and reliable evaluation tools.

Aims: This study aims to examine the quality profile of Arabic PAS items at MAN 1 Gresik by analysing their psychometric characteristics and identifying items that are feasible, need revision, or are not feasible for use.

Methods: This research employed a quantitative descriptive design using psychometric item analysis. The data consisted of 40 multiple-choice PAS items and students' response sheets. The analysis integrated content and construct validity with empirical indicators, including point-biserial validity, KR-20 reliability, item difficulty, and item discrimination, using Microsoft Excel and ANATES V4.

Results: The results show that content validity reached 92.5%, construct validity reached 82.85%, and empirical validity was moderate (r = 0.60). The overall test reliability was high (r_{11} = 0.75). Item difficulty is dominated by medium-level items, while item discrimination is the weakest aspect. Based on integrated psychometric criteria, 40% of items are feasible, 57,5% require revision, and 2,5% are non feasible. The causes of the failure of the test items, content validity (7.5%), construct validity (42.5%), empirical validity (2.5%), level of difficulty (12.5%), and discrimination index (22.5%).

Implications: These findings highlight the importance of systematic psychometric evaluation in Arabic language assessment. Improvements are needed in construct validity, especially Arabic language accuracy, distractor effectiveness, and item discrimination. Such an approach supports the improvement of school-based Arabic assessments to ensure more valid and reliable measurement of students' learning outcomes.

Keywords: Arabic language assessment; content and construct validity; psychometric analysis; final semester test

Article Info:
Received: 16 December 2025
Accepted:27 February 2026
Published: 28 February 2026

How to cite:
Safitri, Z., & Baihaqi, M. (2026). Quality Profile of Arabic Final Semester Items: A Psychometric Analysis. *Al-Lisan: Jurnal Bahasa (e-Journal),11*(1), 87-102. https://doi.org/10.30603/al.v11i1.7322

1. INTRODUCTION

Amidst increasing demands for educational accountability and the need to ensure measurable learning success, the quality of evaluation instruments has become a critical concern in modern education systems (Sibarani et al., 2025; Zahroh & Hilmiyati, 2024). Learning evaluations, such as the Final Semester Assessment (PAS), not only serve as a tool to assess student competency achievement but also serve as a basis for learning decision-making and school quality mapping (Hidayah, 2022; Lam, 2024). In Arabic language learning, assessment of learning outcomes plays a strategic role because it must reflect language skills such as istima', kalam, qiro'ah, and kitabah, as well as linguistic elements such as qawā'id and mufradāt, and because Arabic functions not only as a communication tool but also as a language of religious texts that requires accuracy in meaning and structure, which require representative instruments (Hamid et al., 2022; Umareni et al., 2024). A good instrument must be valid, reliable, and appropriate to the competencies being measured so that the results can be interpreted accurately (Choirudin et al., 2023; Zayrin et al., 2025). However, in practice, the development of PAS in schools is often carried out without systematic review procedures, so that the measured student abilities do not necessarily accurately reflect the targeted competencies (Meyliasari et al., 2024).

This condition reflects broader challenges in school-based assessment practices, particularly in the development of Final Semester Assessments that are not yet supported by systematic quality control procedures. At MAN 1 Gresik, initial reviews indicate that PAS development is conducted as a routine process, but does not include essential evaluation stages such as construct analysis, proportional distribution of cognitive levels, or technical testing of item performance. Although teachers attempt to develop items aligned with instructional materials, limited time, references, and technical support often shift the focus toward content coverage and completion of item numbers rather than item quality (Hayati et al., 2023). Without structured psychometric review, PAS items may become too easy, too difficult, or fail to differentiate student abilities, a pattern that aligns with findings in previous studies (Savika & Zuhriyah, 2024; Yu et al., 2022). This risk directly impacts the interpretation of learning outcomes, enrichment-remedial programs, and other academic decisions (Imran et al., 2025). Therefore, a systematic evaluation of PAS items is necessary to continuously improve the quality of assessments at MAN 1 Gresik.

In this context, psychometric analysis is crucial because it provides a set of procedures for assessing item quality from both a theoretical and judgmental review and empirical perspective. This analysis includes examining content and construct validity, as well as empirical evaluation through empirical validity, reliability, difficulty level, and discriminatory power, allowing for objective identification of the quality of each item (Arbeni et al., 2025; Saputri et al., 2023). Relying solely on teacher or expert reviews tends to yield substantive information. Conversely, using only empirical analysis can neglect the substance of competency (Syafi'i et al., 2025). Psychometric analysis, therefore, integrates both, providing a comprehensive overview of the effectiveness of each test item (İlhan et al., 2024; Savika & Zuhriyah, 2024). This function makes it a highly relevant evaluation tool in the madrasah context, which requires more accurate and consistent assessment instruments. Through this integrative approach, the analysis not only identifies whether test items function statistically, but also explains why certain items fail or succeed in measuring the intended competencies. As a result, psychometric analysis provides actionable feedback for item revision and supports more accurate, data-driven decisions in improving assessment quality.

Although psychometric analysis offers a more comprehensive evaluation framework, this practice has not yet become part of the assessment culture at MAN 1 Gresik. PAS instruments are often considered complete once the test items are developed, without

regular item quality reviews (Thahir, 2023). However, the quality of test items can change due to variations in student characteristics, differences in learning strategies, and curriculum adjustments (Qorib, 2024; Sutomo & Aini, 2024). The absence of data-based feedback makes it difficult for teachers to ensure the consistency of cognitive depth, the accuracy of question construction, and the functionality of each item in differentiating student abilities (Aprilia, 2024; Damayanti et al., 2022; Liani et al., 2025). This study aims to analyze the quality profile of Arabic Language PAS items at MAN 1 Gresik through psychometric analysis that includes content validity, construct validity, empirical validity, reliability, difficulty level, and discriminatory power. This effort is essential to ensure the instrument is truly valid, reliable, and provides an accurate picture of student competency mastery while contributing to improving the quality of learning evaluation and strengthening a professional and accurate assessment culture.

1.1 Research Gap and Novelty

Many studies have been conducted in Indonesia on the evaluation of Arabic language test item quality, but most are partial and do not comprehensively integrate all psychometric aspects. Fikriyah's (2021) study at SMP Muhammadiyah 1 Yogyakarta, for example, only analyzed validity, reliability, difficulty level, and discriminatory power, but did not include construct validity or an examination of cognitive level distribution. Tanjung's (2024) study at MTs Al-Ma'arif Rakit Banjarnegara did include distractor analysis, but found that 70% of the items were invalid without examining technical causes such as cognitive domain inconsistencies or weaknesses in item construction. Harfiani's (2022) study at MAN 2 Kota Bandung focused on analyzing Arabic final semester exam items based on the revised Bloom's taxonomy, primarily examining the cognitive domain without conducting detailed empirical psychometric analysis. Meanwhile, Nizary's (2021) study only focused on content and construct validity of assessment instruments without empirical testing such as empirical validity, reliability, difficulty level, or discriminatory power. Previous findings indicate that existing studies have not integrated all psychometric aspects into a single, comprehensive analysis. Final semester assessments are rarely analyzed substantively and empirically simultaneously. Furthermore, no research has examined the quality of Arabic final semester assessment items at MAN 1 Gresik, even though the Arabic final semester assessment (PAS) is an important basis for determining learning outcomes. The novelty of this research lies in the integration of logical or judgmental review (content and construct validity) with empirical psychometric analysis, including empirical validity, reliability, item difficulty, and item discrimination, within a single analytical framework. Through this integration, the study not only evaluates how items perform statistically, but also explains how substantive alignment influences empirical item performance, thereby providing a clearer basis for item revision and assessment improvement. This research can be used as an evaluation model for teachers and other educational institutions.

1.2 Research Questions

This study seeks to address the following research questions:
1. What is the quality profile of the Arabic Language PAS test items at MAN 1 Gresik based on psychometric indicators?
2. Which items are considered feasible, need revision, or non-feasible based on the results of the integrated psychometric analysis?

2. METHODS

2.1 Research Design

This study employed a quantitative descriptive design using psychometric item analysis to examine the quality profile of Arabic Language Final Semester Assessment (PAS) items for grade XII at MAN 1 Gresik. This design was selected because the research objective was not to test hypotheses, but to describe and map the quality characteristics of each test item based on multiple psychometric indicators. Psychometric item analysis enables systematic evaluation of test items through the integration of logical validation (content and construct validity) and empirical analysis, including empirical validity, test reliability, item difficulty, and discrimination index (Saputra et al., 2022; Sekaran & Bougie, 2016). Each indicator provides specific information about item performance, while their integration allows the researcher to construct a comprehensive quality profile of the assessment instrument. Through this approach, PAS items were not only evaluated individually but also classified into quality categories (feasible, revise, and infeasible) based on integrated psychometric criteria. Therefore, the quantitative descriptive psychometric approach is appropriate for generating a structured and data-based profile of item quality that supports evaluation and improvement of Arabic language assessment instruments.

2.1 Research Objects

The object of this research is the Arabic Language Final Assessment (PAS) questions for grade XII of MAN 1 Gresik in the 2023/2024 academic year. The analysis focuses on multiple-choice questions designed to measure Arabic language skills and linguistic elements. Each item is treated as a unit of analysis analyzed using psychometric criteria, including content validity, construct validity, empirical validity, reliability, difficulty level, and discriminating power. Student answers are used solely as empirical data to estimate item parameters.

2.2 Research Procedures

The research was conducted through five systematic steps as follows:

a) Document Collection
   The researchers collected PAS test scripts, outlines, answer keys, lesson plans, and 30 student answer sheets.
b) Content and Construct Review
   The questions were analyzed qualitatively by subject matter experts and evaluation experts to assess the suitability of indicators, material, cognitive domains, sentence construction, and test writing rules.
c) Scoring and Data Entry
   Students' answers were scored using a dichotomous system (1 = correct, 0 = incorrect). The data were entered into Excel and ANATES for statistical analysis.
d) Empirical Analysis
   Each test item was analyzed using classical test theory techniques, including empirical validity (point-biserial correlation), KR-20 reliability, item difficulty index, and discrimination index. Item difficulty was calculated based on the proportion of students answering each item correctly and classified into three categories: difficult ($p < 0.30$), moderate ($0.30 \leq p \leq 0.70$), and easy ($p > 0.70$). The discrimination index was calculated using the upper–lower group method and interpreted using the

following criteria: poor (D < 0.20), sufficient (0.20 ≤ D < 0.40), good (0.40 ≤ D < 0.70), and very good (D ≥ 0.70). All statistical analyses were conducted using Microsoft Excel and ANATES V4.

e) Conclusion and Categorization
The results of content validity, construct validity, and empirical analysis were integrated to determine the quality status of each item. An item was categorized as feasible if it met content and construct validity criteria and showed feasible empirical performance (valid point-biserial correlation, moderate difficulty, and sufficient or higher discrimination index). Items that met some but not all criteria such as having feasible content relevance but weak empirical indicators were classified as require revision. Items that failed both logical validation and key empirical indicators were categorized as non feasible. This integrative categorization ensured that item quality decisions were based not on a single indicator, but on the combined interpretation of substantive and statistical evidence.

2.3 Research Instruments

The research instruments consisted of three main groups. First, the Arabic Language PAS test sheet consisted of 40 multiple-choice items that were analyzed for content validity, construct validity, and statistical characteristics of the items. Substantive assessment was conducted using the Content and Construct Validity Review Rubric developed based on BSNP standards, covering the suitability of indicators to material, cognitive domain accuracy, item construction, option effectiveness, and language rules. Second, teacher interview guidelines were used as a supporting instrument to obtain information regarding the PAS preparation process, including the basis for question preparation, the absence of standard regulations from the school, and teachers' limitations in conducting item analysis, which has so far only focused on the level of difficulty. Third, Microsoft Excel and ANATES V4 software were used as empirical analysis instruments to calculate empirical validity, reliability, level of difficulty, and discriminatory power so that item quality assessment could be carried out in a standardized and objective manner.

2.4 Data Analysis

Data analysis in this study was conducted through two main stages: logical analysis and empirical analysis, which were systematically formulated according to a psychometric evaluation approach. In the first stage, logical analysis was conducted by examining each item as a stand-alone unit of analysis. Each item was assessed using a review rubric that included the suitability of indicators and material, cognitive domain accuracy, question main construction, option effectiveness, and language rules. This process resulted in content validity and construct validity categories based on the percentage of item feasibility. The second stage, empirical analysis, was conducted using 30 student responses as the basis for calculating the item's statistical characteristics. The analysis includes: (1) empirical validity using the point-biserial correlation coefficient with a minimum limit of $r \geq 0.20$, (2) internal reliability through the r_{11} coefficient with a reliable criterion of $r_{11} \geq 0.70$, (3) the level of difficulty (P) classified into difficult, moderate, and easy, and (4) the discriminating power (D) categorized into poor, good enough, good, and very good. Calculations were performed using Microsoft Excel and ANATES V4 to ensure numerical accuracy and consistency. All logical and empirical analysis results were then integrated to determine the final quality of each item, namely feasible, revised, or not feasible. This integration provides a comprehensive mapping of item performance, thus supporting the drawing of comprehensive conclusions regarding the overall quality of the Arabic Language PAS instrument.

3. FINDINGS AND DISCUSSION

3.1 Findings

The findings of this study are presented as direct answers to two research questions: (1) the quality profile of the Arabic Language Final Semester Assessment (PAS) items at MAN 1 Gresik based on psychometric indicators, and (2) the feasibility category of each item based on the integration of logical and empirical analysis.

Psychometric Quality Profile of Arabic PAS Items

The psychometric quality profile was analyzed through two complementary approaches: logical analysis (content validity and construct validity) and empirical analysis (empirical validity, reliability, item difficulty, and discrimination index). Together, these indicators provide a comprehensive description of how the PAS instrument performs both substantively and statistically.

Content validity analysis was conducted to assess the item's suitability to the indicators, the option structure, and the accuracy of the correct answers. The analysis results indicate that most items met the indicators, although there were some discrepancies, particularly in the cognitive domain. The strongest component was the accuracy of the correct answers, while the weakness lay in the unequal distribution of cognitive levels.

Table 1 Content Validity

Aspects	**Number of items**			
	Valid	**Percentage**	**Invalid**	**Percentage**
Items are aligned with the indicators	37	92,5%	3	7,5%
Answer options are homogeneous and logical	39	97,5%	1	2,5%
Each item has only one correct answer	40	100%	0	0%
Average	**38,6**	**96,6%**	**1.3**	**3.3%**
Items cognitive domains:	**Total**		**Percentage**	
C1	24		60%	
C2	9		22,5%	
C3	7		17,5%	
C4	-		-	
C5	-		-	

The review results showed that 92.5% of the items met the KI-KD indicators, thus substantially representing the competencies taught. However, when viewed from the cognitive domain, the distribution was uneven. The majority of items fell at level C1 (remembering), at 60%, indicating that the instrument emphasized recall rather than understanding or application. The proportions of C2 and C3 items remained, but were not evenly distributed.

Construct validity was analyzed based on 14 indicators reflecting the technical and linguistic quality of the items. The results indicated that several aspects met standards, such as clarity of the question topic, the absence of answer clues, and the use of communicative language. However, significant weaknesses existed in the appropriateness of the Arabic language and the equivalence of the option lengths.

Table 2 Construct Validity

Aspects	Number of items			
	Valid	**Percentage**	**Invalid**	**Percentage**
The statement or stem of the item is formulated clearly and precisely	32	80%	8	20%
The statement or stem does not provide clues to the correct answer	39	97,5%	1	2,5%
The item stem contains only essential statements	40	100%	0	0%
The answer options contain only essential statements	40	100%	0	0%
The item does not depend on the answer to the previous	40	100%	0	0%
Images, graphs, tables, diagrams, or similar materials are presented clearly and are readable (if applicable)	40	100%	0	0%
The final option does not use statements such as "All of the above are correct/incorrect"	40	100%	0	0%
The length of the answer options is relatively similar	29	72,5%	11	27,5%
The item stem is free from double negative statements	40	100%	0	0%
The answer options are arranged in order from the smallest to the largest (based on length of statements, numbers, time sequence, or chronology)	12	30%	28	70%
Each item uses language that conforms to Arabic language rules	0	0%	40	100%
The language used is communicative and easy for students to understand	40	100%	0	0%
The item does not use local or taboo language	40	100%	0	0%
The answer options do not repeat the same words or phrases, unless they form a single unit of meaning	32	80%	8	20%
Average	**33,14**	**82,85%**	**6,85**	**17,14%**

The recapitulation results showed that 82.85% of the items met the item construction standards. However, two major weaknesses emerged: (1) the Arabic language did not follow correct linguistic rules (100% were inconsistent), and (2) the order of the options was inconsistent (70% were inconsistent). These two aspects are significant weaknesses in Arabic language instruments, as linguistic inaccuracies can affect the meaning of the items, and the irregularity of the options can impair readability and increase guessing bias.

Empirical validity was analyzed using two approaches. First, the validity of each item was assessed using the point biserial correlation index between the item score and the

total score by Millman and Greene. Second, ANATES provided the overall test validity results. The minimum validity threshold used was 0.20, in accordance with item analysis standards for multiple-choice tests. The analysis results showed that most items had a positive and adequate correlation with the total score, thus being considered empirically valid. However, one item did not meet the minimum threshold and was categorized as invalid because it failed to demonstrate a strong relationship with the construct being measured.

Table 3 Empirical Validity

Category	Item Numbers	Total	Percentage
Valid	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40	39	97,5%
Invalid	1	1	2,5%

The table shows that 39 items (97.5%) are valid, while 1 item (2.5%) is invalid. Invalid items are generally caused by extreme difficulty, low discrimination, or inconsistent student response patterns. Furthermore, the ANATES results indicate that the overall test validity reached 0.60, which is in the sufficient category. Thus, the PAS instrument generally has adequate empirical validity, although revisions to invalid items are still needed to improve its measurement accuracy.

Reliability was analyzed using the Kuder Richardson 20 (KR-20) because all items are scored dichotomously (1 = correct, 0 = incorrect). The KR-20 analysis produces an internal reliability coefficient that indicates consistency between items in measuring student ability. The criterion used is that an instrument is considered reliable if the KR-20 coefficient is ≥ 0.70, while a value below this limit indicates unreliability (Allen & Yen, 1979).

Table 4 Reliability

Reliability Method	r_{11} Value	Category
KR-20	0,75	Reliable

The table shows that the reliability value of 0.75 indicates that the PAS instrument has high internal consistency, ensuring that the items consistently measure student ability. Therefore, this instrument can be considered reliable and suitable for use in assessing competency achievement in Arabic.

The difficulty level was analyzed to determine the distribution of items in the easy, medium, and difficult categories. Difficulty categories were determined based on the P index, with the following criteria: 0.00–0.30 (difficult), 0.31–0.70 (medium), and 0.71–1.00 (easy) (Zainul & Nasoetion, 1997). The analysis showed that the majority of items fell into the medium category, although there were still relatively large proportions of easy and difficult items.

Table 5 Item Difficulty

Category	Item Numbers	Total	Percentage
Easy	1, 2, 4, 5, 6, 7, 8, 9, 10, 12, 13, 21	12	30%
Moderate	3, 11, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 32, 34, 35, 36, 39	23	57,5%
Difficult	31, 33, 37, 38, 40	5	12,5%

The table shows that 12 items (30%) are classified as easy, 23 items (57.5%) are moderate, and 5 items (12.5%) are difficult. Overall, this distribution is quite proportional, as the majority of items fall into the moderate category. However, the 30% percentage of easy items indicates that some items are too simple and therefore less than optimal in differentiating student abilities. The dominance of easy items also has the potential to reduce discriminatory power and create an imbalance in difficulty. Therefore, several easy items need to be reviewed to increase their difficulty level to meet competency requirements.

Discriminatory power was analyzed to identify the item's ability to differentiate between high-ability and low-ability students (Downing & Haladyna, 2006). Discriminatory power categories were determined based on the D index, with the following criteria: 0.00–0.20 (poor), 0.21–0.40 (adequate), 0.41–0.70 (good), and 0.71–1.00 (very good). The results of the analysis show that most of the items are in the sufficient category, but there are still a number of items with poor discriminatory power so that they do not function optimally in the ability selection procedure.

Table 6 Item Discrimination Index

Category	Item Numbers	Total	Percentage
Poor	1, 3, 8, 12, 20, 29, 31, 37, 39	9	22,5%
Fair	2, 5, 6, 7, 9, 10, 11, 13, 14, 15, 16, 21, 22, 23, 28, 30, 32, 33, 36, 38, 40	21	52,5%
Good	4, 17, 18, 19, 24, 25, 26, 27, 34, 35	10	25%
Very Good	–	0	0%

The table shows that 21 items (52.5%) have sufficient discriminatory power, 10 items (25%) are in the good category, and 9 items (22.5%) have poor discriminatory power. The absence of items with excellent discriminatory power indicates that no questions are truly effective in maximizing differences in ability between students. This condition indicates that the instrument is still not sensitive enough in selecting variations in ability, especially because the proportion of items with poor discriminatory power is relatively high. Therefore, several weak items need to be revised to improve the instrument's discriminatory power.

Integrated Feasibility Categorization of PAS Items

The results of the logical and empirical analyzes were then integrated to produce a final feasibility category for each item. This integrative approach combines content validity, construct validity, empirical validity, difficulty level, and discrimination index, positioning item quality as the outcome of interactions among indicators rather than as isolated measurements. The integration process results in three final categories: Feasible, Revise, and Non-feasible.

Table 7 Integrated Item Categorization

Content Validity	Construct Validity	Empirical Validity	Difficulty Level	Discrimination Index	Final Category
Valid	Valid	Invalid	Easy	Poor	Non-feasible
Valid	Valid	Valid	Easy	Fair	Feasible
Valid	Invalid	Valid	Moderate	Poor	Revise
Valid	Valid	Valid	Easy	Good	Feasible
Valid	Invalid	Valid	Easy	Fair	Revise
Valid	Valid	Valid	Easy	Fair	Feasible
Valid	Valid	Valid	Easy	Fair	Feasible

Valid	Invalid	Valid	Easy	Poor	Revise
Valid	Invalid	Valid	Easy	Fair	Revise
Valid	Valid	Valid	Easy	Fair	Feasible
Valid	Valid	Valid	Moderate	Fair	Feasible
Valid	Valid	Valid	Easy	Poor	Revise
Valid	Valid	Valid	Easy	Fair	Feasible
Valid	Invalid	Valid	Moderate	Fair	Revise
Valid	Invalid	Valid	Moderate	Fair	Revise
Valid	Valid	Valid	Moderate	Fair	Feasible
Valid	Valid	Valid	Moderate	Good	Feasible
Valid	Invalid	Valid	Moderate	Good	Revise
Valid	Valid	Valid	Moderate	Good	Feasible
Valid	Valid	Valid	Moderate	Poor	Revise
Valid	Invalid	Valid	Easy	Fair	Revise
Valid	Valid	Valid	Moderate	Fair	Feasible
Valid	Invalid	Valid	Moderate	Fair	Revise
Valid	Valid	Valid	Moderate	Good	Feasible
Valid	Valid	Valid	Moderate	Good	Feasible
Invalid	Invalid	Valid	Moderate	Good	Revise
Valid	Valid	Valid	Moderate	Good	Feasible
Valid	Invalid	Valid	Moderate	Fair	Revise
Valid	Valid	Valid	Moderate	Poor	Revise
Invalid	Invalid	Valid	Moderate	Fair	Revise
Valid	Valid	Valid	Difficult	Poor	Revise
Valid	Invalid	Valid	Moderate	Fair	Revise
Valid	Invalid	Valid	Difficult	Fair	Revise
Valid	Valid	Valid	Moderate	Good	Feasible
Valid	Valid	Valid	Moderate	Good	Feasible
Invalid	Invalid	Valid	Moderate	Fair	Revise
Valid	Invalid	Valid	Difficult	Poor	Revise
Valid	Invalid	Valid	Difficult	Fair	Revise
Valid	Valid	Valid	Moderate	Poor	Revise
Valid	Valid	Valid	Difficult	Fair	Revise

The findings reveal clear patterns across indicators. Items with poor discrimination index tend to occur at extreme difficulty levels, either easy or difficult (e.g., Items 1, 31, and 37), confirming classical test theory assumptions that items that are too easy or too difficult have limited discriminatory function. In addition, several items demonstrated adequate empirical validity but were categorized as Revise due to weaknesses in construct validity, indicating that empirical performance alone is insufficient to ensure item feasibility. Overall, the integration shows that although the instrument is empirically reliable, the most weaknesses are concentrated in construct-related aspects and discriminatory power, resulting in a dominance of items requiring revision rather than immediate use.

This integrative decision-making process produced a coherent quality profile of the Arabic Language Final Semester Assessment (PAS) instrument. Only a small proportion of items were found to be feasible for direct use without revision, while the majority require improvements. Several items were identified as infeasible due to empirical invalidity or consistently poor discrimination index. This integrative profile highlights the practical contribution of the study by providing teachers with a clear, evidence-based framework for revising and developing higher-quality Arabic language assessment instruments.

Integrated Item Quality Profile

The primary contribution of this study lies in the development of an integrated item quality profile. The dominance of the "revise" category indicates that the Arabic PAS instrument is generally not yet suitable for reuse without improvement, but it cannot be classified as entirely ineffective. Most items still (57.5%) still require revision, while (40%) are feasible, and only 1 item (2.5%) was declared non feasible due to weaknesses in both logical and empirical aspects.

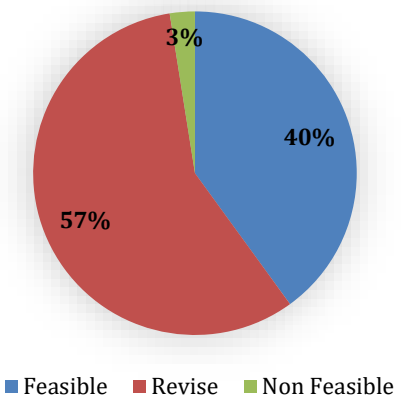

Figure 1. Feasibility of Items

The causes of the failure of the test items are reviewed from 5 indicators, including content validity (7.5%), construct validity (42.5%), empirical validity (2.5%), level of difficulty (12.5%), and discrimination index (22.5%).

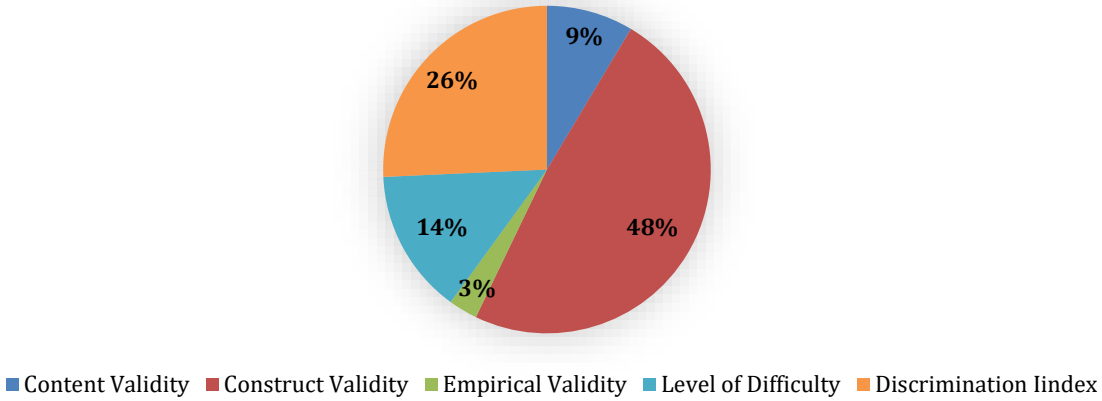

Figure 2. Causes of Items Failure

This profile provides clear guidance for teachers in making item-level decisions regarding retention, revision, or elimination. The findings further suggest that improving the quality of the Arabic PAS instrument does not primarily depend on increasing the number of items or altering test formats, but rather on enhancing teachers' competencies in item construction especially linguistic accuracy, cognitive-level alignment, and distractor effectiveness. Integrating psychometric analysis into school assessment practices can support teachers in shifting from intuition-based item writing toward data-driven decision making. Through the integration of logical and empirical analyses, this study offers an applicable and contextually relevant model for item evaluation in madrasah settings. This approach ensures that assessment instruments are not only statistically reliable but also linguistically accurate and pedagogically meaningful in measuring students' Arabic language competence.

3.2 Discussion

In learning evaluation, the quality of an assessment instrument determines not only its alignment with curricular objectives but also its effectiveness in capturing meaningful differences in students' abilities. In this context, instrument quality cannot be understood merely through individual psychometric indicators, but through how those indicators interact in practice. Therefore, this discussion focuses on the key patterns emerging from the analysis of the Arabic Language PAS instrument, particularly the interdependence among validity, difficulty level, and discrimination index, as well as their implications for item feasibility. the discussion emphasizes why certain items function well while others require revision or elimination, and what these patterns reveal about the strengths and weaknesses of item construction in Arabic language assessment.

Relationships among Psychometric Indicators

The quality of test items is not determined by a single indicator but by the interconnection among content validity, construct validity, empirical validity, item difficulty, and discrimination index, collectively referred to as the integration of psychometric indicators. Items that adequately meet content and construct validity requirements tend to exhibit moderate difficulty levels and higher discrimination indices, enabling them to function effectively in differentiating students' abilities (Hayati et al., 2023).

Conversely, several items displaying extreme difficulty levels and low discrimination indices can be traced to weaknesses in content or construct validity (Nurhasanah et al., 2024). In some cases, item demands are misaligned with competency indicators, causing item difficulty to stem not from higher-order cognitive complexity but from inappropriate content coverage. In other cases, extreme difficulty arises from construct-related issues, such as incomplete stems, non-standard Arabic sentence structures, or distractors that are not syntactically and semantically parallel. Moreover, an empirically "valid" level of validity does not necessarily correspond to optimal discriminatory function when construct validity is weak.

A similar pattern is observed in both very easy and very difficult items. Items that merely require recognition of factual or definitional information generate minimal response variation, while overly difficult items resulting from ambiguous stimuli or non-standard language cause both high-ability and low-ability students to experience comparable difficulty. In both conditions, extreme ease or difficulty directly leads to low discrimination index, even when the item content remains aligned with the indicator (Ebel & Frisbie, 1991). Overall, these findings confirm that relationships among quality indicators are interdependent. Content and construct validity determine the clarity and appropriateness of item demands, whereas empirical validity, item difficulty, and discrimination index reflect how items function in actual testing situations. Therefore, item quality evaluation must be conducted integrative, as weaknesses in one indicator may undermine the performance of others and ultimately affect overall instrument quality.

Construct Validity Problems in Test Design

Within this interrelated psychometric framework, construct validity emerges as the most prominent issue affecting item quality. Although empirical analysis indicates that most items demonstrate feasible empirical validity and the overall test reliability is classified as good (KR-20 = 0.75), the integrated analysis reveals that construct validity weaknesses are the most dominant problem. These weaknesses include inaccurate Arabic sentence structures, inconsistencies in answer option formats, unclear item stems, and irregular distractor patterns.

For example, Item 5 (“تُؤَدِّي مُمَارَسَةُ الرِّيَاضَةِ إِلَى صِحَّةٍ …”) presents multiple options representing different *i'rāb* forms of the word *badaniyyah* (بدنية), yet the stem provides no grammatical cues (such as preceding harakat, syntactic function, or a complete sentence structure). This ambiguity encourages guessing rather than measuring students' ability to apply noun-pattern rules in context. Similarly, Item 3 (“المنعوتُ فِي هَذِهِ الْجُمْلَةِ …”) includes four answer options of relatively similar length; however, the first option (“ظَهَرَ”) is a verb (fi'il) and noticeably shorter, making it a conspicuous and implausible distractor.

In Arabic language assessment, linguistic accuracy is inseparable from the construct being measured (Nurzahira et al., 2025). The finding that all items contain linguistic inaccuracies or imprecision demonstrates that favorable statistical performance does not necessarily indicate adequate construct quality. This explains why several empirically valid items were still categorized as requiring revision. Consequently, this study reinforces the importance of complementing statistical analysis with careful construct and linguistic review, particularly in foreign language assessment.

Distribution of Cognitive Demands

Beyond psychometric and construct considerations, the quality profile of the PAS instrument is also reflected in the distribution of cognitive demands across test items (Sari et al., 2025). One of the major findings of this study is the dominance of low-level cognitive items (C1–C2), which account for more than 80% of the entire test. Although these items are substantively aligned with instructional indicators and learning materials, this distribution highlights the instrument's limited capacity to assess students' intermediate and higher-order thinking skills.

For instance, Item 1 merely requires students to recognize the meaning of the word al-mal'ab (الْمَلْعَبُ) within a clear context, categorizing it as C1. Likewise, Item 3, which asks students to identify na't–man'ūt in a sentence, remains at the level of element recognition (C1). An example of C2 is Item 25, which requires understanding advisory context (“إِلَّا …”) to determine the appropriate meaning. In contrast, items that genuinely assess C3-level skills (application of grammatical rules) are very limited in number.

At the senior secondary level, Arabic language assessment should not merely evaluate memorization of vocabulary or grammatical rules but also students' ability to apply, contextualize, and analyze linguistic structures. The scarcity of C3 items and the absence of C4–C5 items indicate that the assessment remains oriented toward surface-level knowledge (Permendikbud, 2013). This condition may influence classroom instruction, as assessments emphasizing recall tend to encourage similar learning practices.

4. CONCLUSION

The findings indicate that the Arabic Language Final Semester Assessment (PAS) at MAN 1 Gresik has not yet achieved optimal psychometric quality, although several indicators demonstrate adequate performance. Content validity reached 92.5%, and empirical validity showed that 97.5% of the items were valid, supported by a high reliability coefficient (r_{11} = 0.75), indicating that the instrument functions consistently at the measurement level. However, construct validity emerged as the most critical weakness, with an average suitability of 82.85%, particularly in Arabic language accuracy and the consistency of answer option construction, which directly affects item clarity and effectiveness. Difficulty analysis revealed a dominance of moderate items (57.5%), followed by easy (30%) and difficult (12.5%) items, indicating an imbalance in cognitive demand distribution. Item discrimination was identified as the weakest psychometric aspect, with only 25% of items classified as good, while 52.5% were sufficient and 22.5% were poor, limiting the instrument's ability to differentiate student

abilities optimally. Based on integrated psychometric criteria, 40% of items were categorized as feasible, 57.5% required revision, and 2.5% were non-feasible. These findings show that improving the quality of the Arabic PAS instrument depends primarily on strengthening teachers' competencies in linguistically accurate, cognitively aligned, and well-constructed items through integrated logical and empirical psychometric analysis, rather than on changes in test format or item quantity. Future research is recommended to involve larger samples and apply more advanced measurement models, such as Item Response Theory, to obtain a more comprehensive understanding of Arabic language assessment quality beyond the classical test theory approach used in this study.

Acknowledgements
The authors would like to express their sincere gratitude to the leadership and teachers at MAN 1 Gresik for their permission and support in conducting this research. Appreciation is also extended to the lecturers who provided constructive feedback throughout the research process. This research would not have been possible without their cooperation and support.

Authors' Contributions
The first author was responsible for the conception, design, data collection, data analysis, interpretation, and manuscript writing. The second author contributed to the supervision, critical review, and academic guidance throughout the research and writing process.

AI Generative Statement
AI tools were used only for language editing and proofreading. The authors take full responsibility for the content of this manuscript.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Cole Publishing. https://books.google.co.id/books?id=cgElAQAAIAAJ

Aprilia, P. (2024). Cara penanganan siswa berkemampuan di atas rata-rata sedang dan rendah. *Journal of Knowledge and Collaboration*, *1*(7), 311–323. https://doi.org/10.59613/6q3akf79

Arbeni, W., Windiani, A., Sihotang, D. S. B., Anggraini, N., Wulandari, S., & Nugroho, A. (2025). Test reliability analysis in educational evaluation: a quantitative approach to consistency and validity. *Holistic Science*, *5*(1), 59–64. https://doi.org/10.56495/hs.v5i1.838

Choirudin, Sugianto, R., Darmayanti, R., & Muhammad, I. (2023). Teacher competence in the preparation of test and non-test instruments. *Journal of Teaching and Learning Mathematics*, *1*(1), 25–32. https://doi.org/10.22219/jtlm.v1i1.27695

Damayanti, A. M., Daryono, & Rayanto, Y. H. (2022). *Evaluasi pembelajaran*. CV Basya Media Utama. https://books.google.co.id/books?id=cM7cEAAAQBAJ&dq

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Lawrence Erlbaum Associates Publishers. https://psycnet.apa.org/record/2006-01815-000

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. 5th Edition, Prentice-Hall, Englewood Cliffs. https://psycnet.apa.org/record/1973-22100-000

Fikriyah, N. (2021). Analisis butir soal ulangan tengah semester mata pelajaran Bahasa Arab kelas VII semester genap SMP Muhammadiyah 1 Yogyakarta tahun ajaran 2019/2020* [Undergraduate thesis, Universitas Muhammadiyah Yogyakarta]. UMY ETD. https://etd.umy.ac.id/id/eprint/3069/

Hamid, M. A., Sutaman, S., Natsir, M., & Salih, I. O. M. (2022). The development of an evaluation instrument for the implementation of the Arabic language curriculum in Islamic high school. *Jurnal Al Bayan: Jurnal Jurusan Pendidikan Bahasa Arab*, *14*(1), 242–257. https://doi.org/10.24042/albayan.v14i1.10303

Harfiani, M. (2022). Analisis butir soal bahasa Arab kelas XII pada Penilaian Akhir Semester (PAS) semester ganjil tahun ajaran 2021/2022 di MAN 2 Kota Bandung berdasarkan Taksonomi Bloom revisi* [Undergraduate thesis, UIN Sunan Gunung Djati Bandung]. https://digilib.uinsgd.ac.id/55693/

Hayati, R., Wijayati, I. W., Nugroho, F. A., Fazriansyah, M. F., Nurdini, Wardoyo, T. H., Evenddy, S. S., Fratiwi, N. J., Edi, S., Hadikusumo, R. A., Nurlely, L., Mahardiyanti, T., Ariantara, R. G., Tandirerung, V. A., Darmo, S. Y., Suminar, I., Pitrianti, S., Lisnasari, S. F., & Talindong, A. (2023). *Asesmen pembelajaran: teori dan praktik*. PT. Sada Kurnia Pustaka. https://books.google.co.id/books?id=XABbEQAAQBAJ

Hidayah, A. (2022). Internal quality assurance system of education in financing standards and assessment standards. *Indonesian Journal of Education (INJOE)*, *1*(3), 291–300. https://felifa.net/index.php/INJOE/article/view/129

İlhan, M., Güler, N., Teker, G. T., & Ergenekon, Ö. (2024). The effects of reverse items on psychometric properties and respondents' scale scores according to different item reversal strategies. *International Journal of Assessment Tools in Education*, *11*(1), 20–38. https://doi.org/10.21449/ijate.1345549

Imran, I., Bismark, B., Adiansyah, A., Munir, A., & Luthfiyah, L. (2025). Tindak lanjut asesmen pada PAI menjadi program remedial dan pengayaan (teknik memberikan umpan balik dan tindak lanjut hasil asesmen). *Pedagogos: Jurnal Pendidikan*, *7*(1), 49–62. https://doi.org/10.33627/https://doi.org/10.33627/gg.v6i2

Lam, T. N. (2024). Enhancing the quality of competency assessment for elementary school students in modern education. *International Research Journal of Management, IT and Social Sciences*, *11*(3), 93–101. https://doi.org/10.21744/irjmis.v10n3.2429

Liani, A. M., Asmaun, & Nasrullah, A. H. (2025). Peran penilaian yang efektif dalam pengambilan keputusan guru di kelas. *Pedagogy: Jurnal Pendidikan Matematika*, *10*(2), 393–409. https://doi.org/10.30605/pedagogy.v10i2.5904

Meyliasari, A. R., Al-Ibrahimy, A. M., Rohmawati, B., Ariyana, D., Erlindasari, D. P., Nurzaliha, D. P., & Malikah, N. (2024). Penyusunan instrumen penilaian afektif di sekolah. *Muaddib: Jurnal Pendidikan Agama Islam*, *2*(2), 430–441.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In *Educational measurement*. American Council on Education. https://psycnet.apa.org/record/1989-97348-008

Nizary, M. A., & Kholik, A. N. (2021). Validitas instrumen assesmen (Analisis validitas isi dan konstruk instrumen asesmen buku pelajaran Al Quran Hadis kelas 6 Madrasah Ibtidaiyah materi Surat Ad Dhuha bab VI). CONTEMPLATE: Jurnal Pendidikan Bahasa Arab, 2(01), 20-35. https://ejournal.iaiqi.ac.id/index.php/contemplate/article/view/49

Nurhasanah, Hidayatullah, Z., & Arif, M. B. S. (2024). Karakteristik instrumen tes literasi digital ditinjau dari validitas isi dan validitas empiris (kecocokan butir dengan model, reliabilitas, serta tingkat kesukaran butir). *Journal of Classroom Action Research*, *6*(4), 916–923. https://doi.org/10.29303/jcar.v6i4.9650

Nurzahira, F., Jayadi, M. I., & Ridlo, U. (2025). Konsep evaluasi pembelajaran bahasa Arab. *Ihya Al-Arabiyah: Jurnal Pendidikan Bahasa Dan Sastra Arab*, *11*(3), 467–484. http://dx.doi.org/10.30821/ihya.v11i3.26379

Permendikbud. (2013). *Peraturan pemerintah republik Indonesia no. 32 tahun 2013 tentang perubahan atas peraturan pemerintah no. 19 tahun 2005 tentang standar nasional pendidikan*. Menteri Pendidikan dan Kebudayaan Republik Indonesia. https://peraturan.bpk.go.id/Home/Details/5364/pp-no-32-tahun-2013

Qorib, M. (2024). Analysis of differentiated instruction as a learning solution in student diversity in inclusive and moderate education. *International Journal Reglement & Society (IJRS)*, *5*(1), 43–55. https://doi.org/10.55357/ijrs.v5i1.452

Saputra, H. D., Purwanto, W., Setiawan, D., Fernandez, D., & Putra, R. (2022). Hasil belajar mahasiswa: analisis butir soal tes. *Edukasi: Jurnal Pendidikan*, *20*(1), 15–27. https://doi.org/10.31571/edukasi.v20i1.3432

Saputri, H. A. S., Zulhijrah, Larasati, N. J., & Shaleh. (2023). Analisis instrumen asesmen: validitas, reliabilitas, tingkat kesukaran dan daya beda butir soal. *Didaktik: Jurnal Ilmiah PGSD STKIP Subang*, *9*(5), 2986–2995. https://doi.org/10.36989/didaktik.v9i5.2268

Sari, N., Ahmad, Manggaberani, A. A., Jusmiana, A., Metianing, D., Solikhin, F., Negara, H. R. P., Silubun, H. C. A., Disnawati, H., Afri, L. E., Santos, M. Dos, Bahriani, M., & Ningsih, T. Z. (2025). *Konstruksi instrumen pendidikan*. CV Ruang Tentor. https://books.google.co.id/books?id=Neg9EQAAQBAJ&redir

Savika, H. I., & Zuhriyah, I. A. (2024). Peran analisis butir soal terhadap kualitas soal, kompetensi guru, dan prestasi belajar peserta didik di sekolah dasar. *Pandu: Jurnal Pendidikan Anak Dan Pendidikan Umum*, *2*(2), 43–51. https://doi.org/10.59966/pandu.v2i2.856

Sekaran, U., & Bougie, R. (2016). *Research methods for business: a skill building approach*. 7th Edition. John Wiley & Sons, Haddington. https://books.google.co.id/books?id=Ko6bCgAAQBAJ

Sibarani, C. G. G. T., Ahsan, J., & Umar, A. T. (2025). *Buku monograf: evaluasi teori dan model*. CV. Merdeka Kreasi Group. https://books.google.co.id/books?id=NGtxEQAAQBAJ

Sutomo, F. G., & Aini, M. R. Q. (2024). Pemahaman karakteristik peserta didik dalam mengoptimalkan pembelajaran. *Jurnal Kajian Penelitian Pendidikan Dan Kebudayaan*, *2*(4), 60–72. https://doi.org/10.59031/jkppk.v2i4.499

Syafi'i, M., Samsudin, M., Abidin, Z., & Basarrudin, M. (2025). Evaluasi pendidikan sebagai dasar pengembangan instrumen penilaian berbasis kompetensi. *Jurnal Akuntansi, Manajemen Dan Ilmu Pendidikan*, *1*(4), 1–12. https://journal.yapakama.com/index.php/JAMED/article/view/299

Tanjung, M. A. H. R., Fahmi, A. A., Rahmanita, F., Habibah, I. F., & 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), 347–367. https://doi.org/10.25217/mantiqutayr.v4i1.4038

Thahir, M. (2023). *Manajemen mutu sekolah*. Indonesia Emas Group. https://books.google.co.id/books?id=wzraEAAAQBAJ

Umareni, Soehardin, U., & Shodikin, E. N. (2024). Evaluasi pembelajaran bahasa Arab kelas 7 di marhalah salafiyah wustho pondok pesantren Islamic centre bin baz putri Yogyakarta. *Ascent: Al-Bahjah Journal of Islamic Education Management*, *2*(1), 27–35. https://doi.org/10.61553/ascent.v2i1.157

Yu, J., Kreijkes, P., & Salmela-Aro, K. (2022). Students' growth mindset: relation to teacher beliefs, teaching practices, and school climate. *Learning and Instruction*, *80*, 101616. https://doi.org/10.1016/j.learninstruc.2022.101616

Zahroh, F. L., & Hilmiyati, F. (2024). Indikator keberhasilan dalam evaluasi program pendidikan. *Edu Cendikia: Jurnal Ilmiah Kependidikan*, *4*(3), 1052–1062. https://doi.org/10.47709/educendikia.v4i03.5049

Zayrin, A. A., Nupus, H., Maizia, K. K., Marsela, S., Hidayatullah, R., & Harmonedi, H. (2025). Analisis instrumen penelitian pendidikan (uji validitas dan reliabilitas instrumen penelitian). *Qosim: Jurnal Pendidikan Sosial & Humaniora 3.2*, *3*(2), 780–789. https://doi.org/10.61104/jq.v3i2.1070