AL-LISAN: JURNAL BAHASA
Publisher: LPPM IAIN Sultan Amai Gorontalo
ISSN: 2442-8965 E-ISSN: 2442-8973
Volume 11, No. 2 August 2026
Journal Homepage: https://journal.iaingorontalo.ac.id/index.php/al

Psychometric Evaluation of the Arabic Language Fluency Assessment (ALFA): Evidence from Classical Test Theory

[1]Dagi Mardhan Agae Agni
(Corresponding Author)
dagimardhan@gmail.com
Department of Arabic Language Education, Faculty of Tarbiyah and Teacher Training, UIN Sunan Ampel Surabaya, Indonesia

[2]M. Baihaqi
baihaqi@uinsa.ac.id
Department of Arabic Language Education, Faculty of Tarbiyah and Teacher Training, UIN Sunan Ampel Surabaya, Indonesia

ABSTRACT

Background: Language assessment is widely used to determine learners' proficiency; however, the accuracy of score interpretation depends on the quality of the measurement instrument.

Aims: This study aims to evaluate the psychometric quality of the Arabic Language Fluency Assessment (ALFA) using Classical Test Theory. The psychometric findings are subsequently interpreted from the perspective of communicative language assessment informed by the CEFR framework.

Methods: A quantitative descriptive-evaluative design was employed using Classical Test Theory (CTT). Data were collected from 92 beginner-level learners and analyzed using item difficulty, discrimination index, reliability (Cronbach's Alpha), and distractor functioning.

Results: The findings show that the test is dominated by easy items, with mean difficulty indices of 0.93 and 0.88, resulting in limited score variability. Consistent with this pattern, most items exhibited poor discrimination, while a large proportion of distractors were non-functioning. Although the reliability coefficients were relatively high (0.885 and 0.843), these values indicate internal consistency rather than overall measurement quality.

Implications: Within the present sample, the ALFA demonstrated limited ability to differentiate learner performance. The psychometric findings also raise the possibility of construct underrepresentation when interpreted from the perspective of communicative language assessment informed by the CEFR framework. The findings support systematic revision of the ALFA instrument and may also inform the improvement of institution-developed Arabic language assessments in comparable educational contexts.

Keywords: Arabic language testing; CEFR; classical test theory; language assessment; psychometric analysis

Article Info:
Received: 2 May 2026
Accepted: 24 August 2026
Published: 26 August 2026
How to cite:
Agni, D. M. A., & Baihaqi, M. (2026). Psychometric evaluation of the Arabic language fluency assessment (ALFA): Evidence from classical test theory. *Al-Lisan: Jurnal Bahasa (e-Journal), 11*(2), 237-250.
https://doi.org/10.30603/al.v11i2.7830

1. INTRODUCTION

Language assessment plays a central role in determining learners' proficiency and informing instructional decisions in language education (Nasr, 2026). However, test scores cannot be interpreted as direct indicators of language ability without considering the quality of the measurement instrument (Girolamo et al., 2022). In contemporary assessment theory, the meaning of test scores depends not only on numerical outcomes but also on the extent to which the instrument validly represents the construct being measured (Sireci & Rodriguez, 2022). Consequently, poorly designed tests may produce results that appear accurate yet fail to reflect learners' actual competence (El Chaal et al., 2025). This highlights the necessity of systematically evaluating the quality of language assessment instruments to ensure that score interpretations are meaningful and defensible (Bond & Fox, 2021).

In many institutional contexts, particularly in Arabic language education, assessment practices are often developed locally and implemented as part of routine evaluation without rigorous validation (Zeinoun et al., 2022). Rigorous validation is especially important in Arabic language assessment because diglossia, dialectal variation, and morphological complexity may affect how language ability is measured (Rakhlin et al., 2021). These institution-based tests are widely used for placement, grading, and certification purposes; however, their psychometric quality is rarely examined through systematic empirical analysis (Subando, 2025). As a result, test scores are frequently treated as accurate representations of learner proficiency, despite the possibility that the underlying test design does not adequately capture the complexity of language ability (Ljubojević & Daničić, 2026; Subando, 2025). This situation highlights a discrepancy between the intended function of assessment and the quality of evidence generated by the instrument.

A major limitation in this context lies in the limited integration of core psychometric analyses, particularly item difficulty, discrimination indices, and reliability estimates, resulting in an incomplete evaluation of test quality (El Chaal et al., 2025). These parameters are fundamental in determining whether a test can meaningfully differentiate between learners of varying ability levels (Sharma, 2021). When such properties are not systematically analyzed, it becomes difficult to establish whether observed scores reflect true language competence or merely the structural characteristics of the test (Subando, 2025). In addition, the role of distractor functioning in multiple-choice items is often overlooked, even though it directly influences response behavior and overall test performance (Brown & Abeywickrama, 2019). The absence of comprehensive item analysis may therefore result in instruments that appear consistent but lack meaningful measurement precision (Zhao & Aryadoust, 2025).

Beyond psychometric concerns, another critical issue lies in the alignment between test design and established frameworks of language proficiency, particularly the Common European Framework of Reference for Languages (CEFR). Within this framework, language ability is conceptualized as the capacity to use language meaningfully in communicative contexts rather than merely recognizing linguistic forms (Council of Europe, 2020). Nevertheless, many assessment instruments continue to emphasize surface-level features such as vocabulary and grammatical recognition, with limited attention to communicative competence (Alamer et al., 2025). This imbalance leads to construct underrepresentation, where essential dimensions of language ability are not adequately captured, thereby undermining the validity of score interpretation (Fulcher, 2010). The CEFR was selected because it provides an internationally recognized framework for describing communicative language proficiency and has increasingly informed the development and interpretation of language assessments across educational contexts (Council of Europe, 2020).

1.1 Research Gap and Novelty

Recent studies have increasingly applied psychometric techniques to evaluate language assessment instruments, although their scope remains fragmented. For example, Sharma (2021) examined item difficulty, discrimination, and distractor efficiency in English speech-sound assessments, demonstrating the importance of item analysis for improving test quality. Similarly, El Chaal et al. (2025) evaluated the psychometric performance of human-crafted and AI-generated multiple-choice questions, focusing on item-level statistics and reliability. In Arabic language assessment, Subando (2025) employed a multidimensional Item Response Theory approach to investigate item characteristics in Madrasah Aliyah examinations, while Zeinoun et al. (2022) reviewed methodological practices in Arabic psychological testing and emphasized the need for more rigorous validation procedures. Collectively, these studies have contributed to either psychometric evaluation or construct-oriented language assessment; however, they have generally treated these perspectives separately. Consequently, limited empirical evidence explains how psychometric findings from institution-developed Arabic language assessments can inform discussions of construct representation in communicative language assessment.

Building upon previous studies, two important gaps remain. First, psychometric evaluation and construct-oriented language assessment have generally been investigated separately, with limited attention to how empirical evidence from item functioning informs the interpretation of language constructs. Second, empirical studies on institution-developed Arabic language assessments remain scarce, particularly those providing comprehensive evidence from item difficulty, discrimination, reliability, and distractor functioning.

This study addresses these gaps by conducting a comprehensive psychometric evaluation of the Arabic Language Fluency Assessment (ALFA) using Classical Test Theory. The findings are subsequently interpreted from the perspective of communicative language assessment informed by the CEFR framework, providing empirical evidence on the psychometric characteristics of an institution-developed Arabic language assessment and offering a construct-oriented interpretation of the findings within the CEFR framework.

1.2 Research Questions

This study is guided by the following research questions:

1. How does the Arabic Language Fluency Assessment (ALFA) perform in terms of its psychometric quality based on Classical Test Theory?
2. What do the psychometric findings imply for communicative language assessment from the perspective of the CEFR framework?

Accordingly, this study aims to evaluate the psychometric quality of the ALFA by examining its item difficulty, discrimination, internal consistency, and distractor functioning using Classical Test Theory. It also aims to interpret these psychometric findings from the perspective of communicative language assessment informed by the CEFR framework, particularly in relation to the extent to which the assessment provides meaningful evidence of beginner-level language proficiency.

2. METHODS

2.1 Research Design

This study employed a quantitative descriptive–evaluative design to investigate the psychometric properties of the Arabic Language Fluency Assessment (ALFA). The quantitative approach was utilized to enable objective measurement of item and test performance based on numerical response data. The descriptive–evaluative orientation

was selected because the study focuses on examining the quality of an existing instrument rather than testing causal relationships.

The analysis was grounded in Classical Test Theory (CTT), which remains a widely accepted framework for preliminary test evaluation, particularly in institutional assessment contexts with moderate sample size. CTT facilitates the estimation of item-level statistics such as difficulty, discrimination, and internal consistency, which are essential for identifying poorly functioning items and overall test quality (Ayanwale et al., 2022).

Although more advanced models such as Item Response Theory (IRT) or Rasch measurement offer stronger parameter invariance, CTT was considered appropriate for this study due to its suitability for initial validation stages, its lower sample size requirements, and its practical applicability in institutional test development settings. Therefore, this design provides a foundational empirical evaluation of the instrument prior to potential further analysis using more sophisticated models.

2.2 Participants

The participants consisted of 92 undergraduate students enrolled in an Arabic language program at STAI Ali bin Abi Thalib Surabaya. All participants were categorized as beginner-level learners based on the institutional curriculum, which corresponds approximately to the A1–A2 levels of the Common European Framework of Reference for Languages (CEFR). At this level, learners are expected to demonstrate basic vocabulary recognition, simple grammatical awareness, and limited comprehension of spoken and written input (Council of Europe, 2020). As the ALFA was designed to assess beginner-level proficiency, this sample appropriately represented the target population. However, the relatively homogeneous proficiency level of the participants may have restricted score variability, a factor that should be considered when interpreting psychometric indices, particularly item discrimination.

A census sampling approach was employed, whereby all students who completed the ALFA test during the designated testing period were included in the dataset. This approach ensured complete coverage of the target population because all eligible participants were included in the analysis, thereby minimizing sampling bias within the institutional context (Devianti et al., 2025).

The sample size was considered sufficient for estimating item difficulty within the Classical Test Theory framework, consistent with recommendations for preliminary educational measurement studies (DeVellis, 2017). However, the discrimination analysis, based on the upper–lower 27% method, involved relatively small comparison groups (approximately 25 participants per group), and the distractor analysis was based on low response frequencies for individual options. Accordingly, the results of the discrimination and distractor analyses should be interpreted with appropriate caution, as these estimates may be less stable than the item difficulty indices.

2.3 Research Procedures

The study utilized response data obtained from the administration of the ALFA test as part of the institution's regular assessment program. The ALFA was administered to 92 undergraduate students enrolled in the Arabic language program during the designated testing period. After the assessment was completed, the students' response data were compiled into a dataset and provided to the researchers for analysis. The dataset contained the responses to all 70 multiple-choice items for each participant, which were subsequently coded and organized into a person-by-item response matrix for psychometric analysis. Participant identities were anonymized prior to analysis to maintain confidentiality.

All responses were coded dichotomously, with correct answers assigned a value of 1 and incorrect answers assigned a value of 0. The coded responses were then organized into a person-by-item response matrix, which serves as the standard input format for psychometric analysis (Bond & Fox, 2021).

Prior to analysis, the dataset was screened to ensure completeness and consistency. Responses with missing values were checked; however, no substantial missing data were identified that required exclusion or imputation. Total scores for each participant were subsequently computed and used as the basis for grouping participants in discrimination analysis (Brown & Abeywickrama, 2019). All statistical analyses were conducted using Excel, ensuring systematic and reproducible data processing procedures.

2.4 Research Instruments

The instrument analyzed in this study was the Arabic Language Fluency Assessment (ALFA), an institution-developed test designed to measure functional language ability at the beginner level. The test was constructed to align with the institution's curriculum, which emphasizes foundational communicative competence, including listening comprehension and basic grammatical recognition.

The ALFA consists of two subtests: *istimā'* (listening comprehension), comprising 40 multiple-choice items, and *tarākīb* (structure and vocabulary), comprising 30 multiple-choice items. Each item includes four response options with one correct answer.

The listening subtest measures the ability to understand simple spoken utterances, including everyday expressions and short informational statements. The structure subtest assesses learners' recognition of basic grammatical patterns and vocabulary usage in controlled contexts. These constructs broadly correspond to beginner-level descriptors outlined in the CEFR (Council of Europe, 2020).

Prior to this study, the instrument had undergone internal expert review to ensure content relevance; however, it had not yet been subjected to empirical psychometric validation. Therefore, this study aims to evaluate the measurement quality of the instrument through statistical analysis.

2.5 Data Analysis

Data analysis was conducted within the framework of Classical Test Theory (CTT) to evaluate the psychometric quality of the ALFA instrument. The analysis focused on four key components: item difficulty, discrimination index, reliability, and distractor functioning (El Chaal et al., 2025). These components should be interpreted together because they provide complementary evidence of item quality (Najmalia Fitra et al., 2025). No formal CEFR alignment analysis, such as descriptor mapping or expert-rating procedures, was undertaken. Instead, the CEFR framework served as a conceptual framework for interpreting the psychometric findings in the Discussion from the perspective of communicative language assessment.

Item difficulty (P) was calculated as the proportion of test-takers who answered each item correctly, with values ranging from 0 to 1 (Yu & Xu, 2024). Items with P values between 0.30 and 0.70 were considered to have acceptable difficulty levels, while values below 0.30 indicated difficult items and values above 0.70 indicated easy items (El Chaal et al., 2025).

The discrimination index (D) was computed using the upper–lower group method. Participants were divided into high and low groups based on the top and bottom 27% of total scores (Ebel & Frisbie, 1991). Because the upper–lower method may be sensitive to skewed score distributions, the discrimination indices were interpreted cautiously. The discrimination index for each item was calculated as the difference in the proportion of correct responses between these two groups. Items with $D \geq 0.30$ were considered good

discriminators, values between 0.20–0.29 were considered acceptable, and values below 0.20 indicated poor discrimination (El Chaal et al., 2025).

Test reliability was estimated using Cronbach's Alpha coefficient to assess internal consistency. A reliability coefficient of α ≥ 0.70 was considered acceptable for research purposes, while values above 0.80 indicated good reliability (Cronbach, 1951).

In addition, distractor analysis was conducted to evaluate the effectiveness of incorrect answer options (Brown & Abeywickrama, 2019). Distractors selected by less than 5% of participants were classified as non-functioning, as they contribute minimally to item discrimination (Rodriguez, 2005).

All statistical results were interpreted systematically to identify strengths and weaknesses of the instrument and to provide recommendations for item revision and test improvement.

3. FINDINGS AND DISCUSSION

3.1 Findings

This section presents the findings in relation to the two research questions. The first research question concerns the psychometric quality of the Arabic Language Fluency Assessment (ALFA) based on Classical Test Theory, including item difficulty, discrimination, reliability, and distractor functioning. The second research question concerns the implications of these psychometric findings for communicative language assessment from the perspective of the CEFR framework. Accordingly, the findings first present the psychometric characteristics of the ALFA and subsequently interpret their implications for communicative language assessment.

The descriptive analysis revealed a notably high level of performance across both subtests of the Arabic Language Fluency Assessment (ALFA). The mean score for the *istimā'* subtest was 37.07 (SD = 3.77) out of a maximum score of 40, corresponding to 92.7%. Similarly, the *tarākīb* subtest yielded a mean score of 27.32 (SD = 5.88) out of 30, equivalent to 91.1%.

Table 1 Descriptive Statistics of ALFA Scores

Subtest	Number of Items	Mean Score	Percentage	Standard Deviation
Istimā'	40	37.07	92.7%	3.77
Tarākīb	30	27.32	91.1%	5.88

These results indicate that the majority of participants achieved scores close to the upper limit of the scale, suggesting a strong concentration of high performance. The relatively small standard deviation in the *istimā'* subtest, in particular, reflects limited variability in participant responses.

The score distribution further demonstrates a clear clustering effect at the upper end of the scale, with fewer observations in the middle and lower score ranges. Such a pattern suggests a potential ceiling effect, where the test may not sufficiently differentiate among higher-ability learners. This tendency is more pronounced in the *istimā'* subtest, where the score range is narrower relative to the maximum possible score.

`

The analysis of item difficulty revealed that the overall difficulty level of the test was relatively low.

Table 2 Distribution of Item Difficulty

Category	***Istimā'***	***Tarākīb***
Easy	38	29
Medium	1	1
Difficult	1	0

The mean difficulty index (P) was 0.93 for the *istimā'* subtest and 0.88 for the *tarākīb* subtest, indicating that most items were answered correctly by a large proportion of participants.

The distribution of item difficulty confirms this pattern. In the *istimā'* subtest, 38 out of 40 items were classified as easy, with only one item categorized as medium and one as difficult. Similarly, in the *tarākīb* subtest, 29 out of 30 items were classified as easy, with only one item falling into the medium category and none categorized as difficult.

Notably, several items in the *istimā'* subtest reached a difficulty index of 1.00, indicating that all participants answered these items correctly. This finding suggests that the majority of test items are insufficiently challenging for the target population.

The analysis of item discrimination revealed a substantial limitation in the ability of the test to distinguish between higher- and lower-performing participants. The results show that a large proportion of items in both subtests fall into the poor discrimination category.

Table 3 Item Discrimination Index

Category	***Istimā'***	***Tarākīb***
Good	5	8
Fair	5	2
Poor	30	20

In the *istimā'* subtest, 30 out of 40 items demonstrated poor discrimination, while only five items were classified as good and five as fair. A similar pattern was observed in the *tarākīb* subtest, where 20 out of 30 items showed poor discrimination. This finding should be interpreted in conjunction with the item difficulty results. Because most items were classified as very easy, with several approaching a difficulty index of 1.00, the low discrimination indices are partly an expected mathematical consequence of the restricted variability in item responses rather than entirely independent evidence of poor item functioning.

This distribution indicates that most items do not effectively differentiate between varying levels of participant ability. The dominance of poorly discriminating items suggests that the test may lack sensitivity in capturing meaningful differences in language proficiency.

The reliability analysis indicates that both subtests demonstrate a high level of internal consistency. The Cronbach's Alpha coefficient was 0.885 for the *istimā'* subtest and 0.843 for the *tarākīb* subtest.

Table 4 Reliability Coefficients of ALFA

Subtest	**Cronbach's Alpha**
Istimā'	0.885

Subtest	Cronbach's Alpha
Tarākīb	0.843

These values exceed the commonly accepted threshold for reliability in educational measurement, suggesting that the items within each subtest demonstrate good internal consistency. However, the reliability coefficients should be interpreted cautiously, as high Cronbach's Alpha reflects the consistency of responses among items rather than the overall quality of the measurement instrument. In the present study, the relatively high alpha values are likely attributable to the substantial number of items and the consistent measurement of a similar construct, whereas the instrument's limited discrimination should be interpreted separately.

The analysis of distractor functioning revealed a pervasive issue across both subtests. A very high proportion of items contain non-functioning distractors, defined as options selected by less than 5% of participants.

Table 5 Distractor Functioning Statistics

Subtest	Total Items	% ≥1 NFD	% ≥2 NFD	% All Distractors Functioning	Mean NFD per Item
Istimā'	40	97.5%	95.0%	2.5%	2.65
Tarākīb	30	100%	80.0%	0%	2.40

In the *istimā'* subtest, 97.5% of items included at least one non-functioning distractor, and 95.0% contained two or more. Only 2.5% of items had all distractors functioning effectively. Similarly, in the *tarākīb* subtest, all items (100%) contained at least one non-functioning distractor, and 80.0% contained two or more. No items were identified in which all distractors functioned effectively.

The mean number of non-functioning distractors per item was 2.65 for *istimā'* and 2.40 for *tarākīb*, indicating that most items effectively operate with only one plausible alternative option. This pattern suggests that distractor quality is generally weak and may contribute to the overall ease of the test and its limited discriminative power.

3.2 Discussion

Integrated Psychometric Evaluation of the ALFA

The consistently high mean scores observed across both subtests should not be interpreted as direct evidence of high language proficiency (Subando, 2025). Within a measurement framework, observed scores are a function of both learner ability and test characteristics (Fulcher, 2010). The clustering of scores at the upper end of the scale indicates a restricted score range, suggesting a ceiling effect (Bahnson et al., 2023). Under such conditions, the test provides limited information at higher ability levels, thereby weakening its capacity to distinguish among more proficient learners (Ljubojević & Daničić, 2026). Consequently, the resulting scores risk producing an illusion of proficiency, where high performance reflects test easiness rather than actual competence (Cardwell et al., 2023). The Qiyas Arabic language test showed that reliable scores may still provide limited coverage of learners' abilities, particularly at the middle and upper levels (Al-Owidha, 2018).

This pattern is closely linked to the dominance of easy items, as reflected in the high difficulty indices (Subando, 2025). When most items are answered correctly, score variance becomes compressed, leading to attenuation in the relationship between observed scores and the underlying latent trait (Embretson & Reise, 2000). From a psychometric perspective, reduced variance diminishes the test's information function,

particularly in the upper ability range, where differentiation is most needed (Bond & Fox, 2021). As a result, the test shifts from a measurement instrument into a confirmation tool of basic knowledge, limiting its interpretive power across the ability continuum (El Chaal et al., 2025).

The prevalence of low discrimination indices further reinforces this limitation. Items with weak discrimination contribute minimally to differentiating ability because they produce similar response patterns across high- and low-performing groups (Brown & Abeywickrama, 2019). This indicates that the items fail to align response probability with ability level, which is a core requirement in both Classical Test Theory and modern measurement models (Embretson & Reise, 2000). Consequently, the test yields low diagnostic value, as it cannot reliably separate learners based on functional proficiency (Subando, 2025).

However, this interpretation should also be considered in light of the relatively homogeneous proficiency level of the participants. Because all examinees represented beginner-level learners and completed a test designed for the same proficiency level, the restricted range of ability may have contributed to the limited score variability and the relatively low discrimination indices observed in this study. Accordingly, the findings should be interpreted as reflecting both the psychometric characteristics of the instrument and the characteristics of the sampled population.

Although both subtests demonstrated relatively high Cronbach's Alpha coefficients, these values should be interpreted with caution (Fulcher, 2010). Cronbach's Alpha reflects the internal consistency of item responses and is influenced by factors such as test length, dimensionality, and average inter-item correlation rather than serving as a direct indicator of overall measurement quality. Consequently, high internal consistency does not necessarily indicate that an instrument effectively differentiates learners with different ability levels, particularly when most items exhibit limited discrimination. These findings illustrate an important psychometric consideration: evidence of good internal consistency can coexist with limited item discrimination because these indices evaluate different aspects of measurement quality. Therefore, reliability estimates should be interpreted alongside other psychometric evidence when evaluating the overall quality of a language assessment instrument (Cronbach, 1951).

CEFR-Informed Interpretation of the Psychometric Findings

From a construct perspective, the psychometric findings may have implications for how the ALFA test represents communicative language ability as conceptualized in the CEFR. The predominance of very easy items, limited item discrimination, and ineffective distractors may indicate that the instrument places greater emphasis on recognition of basic linguistic forms rather than providing strong evidence of broader communicative language ability. Although this study did not conduct a formal CEFR alignment analysis, these findings raise the possibility of construct underrepresentation when interpreted from the perspective of the CEFR framework (Council of Europe, 2020). Consequently, this interpretation should be regarded as a hypothesis generated by the psychometric evidence rather than as an empirically established conclusion. This reflects construct underrepresentation, where essential dimensions of the target construct are insufficiently captured (Haladyna & Rodriguez, 2013). As a result, the test provides limited evidence to support claims about learners' communicative proficiency (Chapelle, 2021).

The analysis of distractor functioning further clarifies this issue. The high proportion of non-functioning distractors reduces item complexity and lowers cognitive demand, effectively transforming items into recognition-based tasks (Brown & Abeywickrama, 2019). Effective distractors can increase item difficulty and discrimination, whereas weak distractors may reduce an item's ability to distinguish between examinees (Ludewig et al., 2023). This weakens the response process by minimizing the need for inferential or strategic thinking, thereby reducing response variability and further limiting

discrimination (Schmucker & Moore, 2026). In this sense, distractor inefficiency contributes directly to the broader measurement problem rather than representing an isolated flaw (Rezigalla et al., 2024).

Importantly, these findings address a notable gap in the literature on Arabic language assessment, where psychometric evaluations of institution-based tests remain relatively limited and are often restricted to descriptive reporting without integrated analysis of item functioning (Subando, 2025). While previous studies have examined aspects of test performance, few have systematically demonstrated how item difficulty, discrimination, reliability, and distractor functioning interact as a unified measurement system (Zeinoun et al., 2022). By situating these components within a coherent psychometric framework, the present study contributes to a more comprehensive understanding of test quality in Arabic language assessment contexts (Subando, 2025).

Taken together, these findings demonstrate that the limitations of the ALFA test are systemic. Item difficulty, discrimination, reliability, and distractor functioning operate as interdependent components within a psychometric system (Rezigalla et al., 2024). The overrepresentation of easy items compresses score variance, which weakens item discrimination and reduces the interpretability of test scores. Ineffective distractors further simplify response behavior, reinforcing uniform response patterns (Haladyna & Rodriguez, 2013). As a result, the test provides insufficient internal evidence to support valid and meaningful score interpretation.

These findings imply that improving the ALFA test requires systematic redesign rather than incremental revision. Test development should be guided by a construct-driven approach, ensuring alignment with communicative language use. A balanced distribution of item difficulty is necessary to maintain measurement precision across the ability spectrum, while items should be designed to elicit varying levels of cognitive processing to enhance discrimination (AlKhuzaey et al., 2024). Distractors should be constructed based on empirically observed learner errors to improve plausibility and effectiveness (Rodriguez, 2005). Moreover, task design should move toward meaning-based processing that reflects authentic communicative demands (Brown & Abeywickrama, 2019).

At a broader level, these results highlight the need to strengthen assessment practices in Arabic language education (Zeinoun et al., 2022). Institutions should prioritize assessment literacy, adopt construct-based frameworks, and integrate psychometric analysis as a routine component of test development (Kremmel & Harding, 2020; Kunnan, 2020). In this way, language assessment can function not only as a grading mechanism but also as a diagnostic tool that generates meaningful insights into learner development (Fulcher & Harding, 2022).

Limitations

This study has several limitations. First, the analysis was based on a single institution and one locally developed assessment instrument (ALFA), which limits the generalizability of the findings to other Arabic language assessment contexts. Second, all participants represented a relatively homogeneous beginner-level population (approximately CEFR A1–A2), and the restricted ability range may have influenced score variability, item discrimination, and distractor performance. Third, the psychometric evaluation was conducted solely within the Classical Test Theory framework; therefore, item parameter invariance and person ability estimates were not examined. Future research should employ larger and more heterogeneous samples and apply modern measurement models, such as Rasch Measurement Theory or Item Response Theory, to provide more robust psychometric evidence. Finally, this study did not include external sources of validity evidence, such as criterion-related validity, response process analysis, or empirical CEFR alignment procedures. Accordingly, the discussion of communicative language assessment should be interpreted as a construct-oriented interpretation rather than as an empirically established alignment analysis, and the findings should be

interpreted within the context of institution-developed beginner-level assessments rather than generalized to Arabic language assessments more broadly.

4. CONCLUSION

This study concludes that, within the present sample, the Arabic Language Fluency Assessment (ALFA) demonstrated limited effectiveness in differentiating learner ability despite showing high internal consistency. The predominance of easy items, limited item discrimination, and the prevalence of non-functioning distractors contributed to restricted score variability and reduced measurement sensitivity. Consequently, the resulting test scores provide limited evidence for supporting robust interpretations of learners' language proficiency within the context of this study.

From a construct perspective, the psychometric findings raise the possibility that the ALFA may place greater emphasis on surface-level linguistic recognition than on broader aspects of communicative language ability as conceptualized in the CEFR. Because no formal CEFR alignment analysis was conducted, this interpretation should be regarded as a construct-oriented hypothesis rather than an empirically established conclusion.

These findings highlight the need for systematic revision of the Arabic Language Fluency Assessment (ALFA), particularly with respect to item difficulty, item discrimination, distractor quality, and construct representation. Similar considerations may also be applicable to institution-developed Arabic language assessments implemented in comparable educational contexts. Future revisions of the ALFA should prioritize a more balanced distribution of item difficulty, improved item discrimination, and more effective distractor design to strengthen the interpretability of test scores. Future studies involving larger and more heterogeneous samples, together with modern psychometric approaches such as Rasch Measurement Theory or Item Response Theory, are recommended to further validate and refine institution-developed Arabic language assessments.

Acknowledgements

The authors express their sincere appreciation to the management and academic staff of STAI Ali bin Abi Thalib Surabaya for their support and cooperation throughout the research process. The authors also acknowledge the students who participated in the administration of the Arabic Language Fluency Assessment (ALFA) and contributed the response data used in this study. Their participation and cooperation were essential to the completion of this research. The authors further appreciate the constructive feedback provided by colleagues and reviewers, which contributed to strengthening the analytical rigor and clarity of this manuscript.

Authors' Contribution

Dagi Mardhan Agae Agni conceptualized the study, formulated the research design, conducted the data collection and analysis, interpreted the findings, and prepared the initial and revised versions of the manuscript. M. Baihaqi supervised the research process, provided guidance on the research design, methodology, interpretation of findings, and manuscript development, and critically reviewed the manuscript. Both authors approved the final version of the manuscript and agreed to be accountable for the integrity and scholarly rigor of the work.

AI Generative Statement

The authors declare that artificial intelligence tools were used in a limited capacity to support language refinement, grammar checking, and clarity of expression during manuscript preparation. All research conceptualization, data collection, data analysis, interpretation of findings, and final academic decisions were conducted independently

by the authors. The authors take full responsibility for the originality, accuracy, interpretation, and integrity of the content presented in this article.

REFERENCES

Alamer, A., Al Khateeb, A., & Alshabeb, A. (2025). The creation and validation of the Arabic vocabulary levels Test (Arabic-VLT). *Language Assessment Quarterly*, *22*(2), 117–137. https://doi.org/10.1080/15434303.2025.2477445

AlKhuzaey, S., Grasso, F., Payne, T. R., & Tamma, V. (2024). Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, *34*(3), 862–914. https://doi.org/10.1007/s40593-023-00362-1

Al-Owidha, A. A. (2018). Investigating the psychometric properties of the qiyas for L1 Arabic language test using a rasch measurement framework. *Language Testing in Asia*, *8*(1), 12. https://doi.org/10.1186/s40468-018-0064-5

Ayanwale, M. A., Chere-Masopha, J., & Morena, M. C. (2022). The classical test or item response measurement theory: The status of the framework at the examination council of Lesotho. *International Journal of Learning, Teaching and Educational Research*, *21*(8), 384–406. https://doi.org/10.26803/ijlter.21.8.22

Bahnson, M., Sallai, G., Jwa, K., & Berdanier, C. G. P. (2023). Mitigating ceiling effects in a longitudinal study of doctoral engineering student stress and persistence. *International Journal of Doctoral Studies*, *18*, 199–227. https://doi.org/10.28945/5118

Bond, T. G., & Fox, C. M. (2021). *Applying the rasch model: Fundamental measurement in the human sciences*. Routledge. https://doi.org/10.4324/9781410614575

Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices*. Pearson.

Cardwell, R., Naismith, B., LaFlair, G. T., & Nydick, S. (2023). Duolingo English test: Technical manual. *Duolingo Research Report*, 1–33.

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. SAGE Publications. https://doi.org/10.4135/9781071878811

Council of Europe. (2020). *Common european framework of reference for languages: Learning, teaching, assessment*. Council of Europe Publishing.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

DeVellis, R. F. (2017). *Scale development: Theory and applications*. SAGE Publications.

Devianti, R., Subiyanto, D., & Ratna Purnamarini, T. (2025). Digital transformation as a driver of innovative behavior: The mediating roles of hr analytics and psychological well-being at the library and archives office of Bantul regency. *International Journal of Economics and Management Review*, *3*(3), 46–60. https://doi.org/10.58765/ijemr.v3i3.275

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall.

El Chaal, R., Seghir, R. Ben, & Aboutafail, M. O. (2025). Psychometric evaluation of human-crafted and ai-generated multiple-choice questions for Mathematics instruction. *Education Science and Management*, *3*(2), 78–92. https://doi.org/10.56578/esm030201

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Fulcher, G. (2010). *Practical language testing*. Hodder Education. https://doi.org/10.4324/9780203767399

Fulcher, G., & Harding, L. (2022). *The Routledge handbook of language testing* (2nd ed.). Routledge. https://doi.org/10.1080/13803611.2023.2179073

Girolamo, T., Ghali, S., Campos, I., & Ford, A. (2022). Interpretation and use of standardized language assessments for diverse school-age individuals. *Perspectives of the ASHA Special Interest Groups*, *7*(4), 981–994. https://doi.org/10.1044/2022_PERSP-21-00322

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge. https://doi.org/10.4324/9780203850381

Kremmel, B., & Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. *Language Assessment Quarterly*, *17*(1), 100–120. https://doi.org/10.1080/15434303.2019.1674855

Kunnan, A. J. (2020). Evaluating language assessments from an ethics perspective. *JLTA Journal*, *23*, 3–13. https://doi.org/10.20622/jltajournal.23.0_3

Ljubojević, D., & Daničić, M. (2026). Assessing oral proficiency in young EFL learners: A study of eighth-grade pupils' performance at the serbian national English competition. *Research in Pedagogy*, *16*(1), 22–35. https://doi.org/10.5937/IstrPed2601022L

Ludewig, U., Schwerter, J., & McElvany, N. (2023). The features of plausible but incorrect options: Distractor plausibility in synonym-based vocabulary tests. *Journal of Psychoeducational Assessment*, *41*(7). https://doi.org/10.17877/DE290R-24398

Najmalia Fitra, Herdah, H., & Mahrous, A. E. (2025). Language test item analysis techniques: Teknik analisis item tes bahasa. *Al Mahāra: Jurnal Pendidikan Bahasa Arab*, *11*(1), 158–179. https://doi.org/10.14421/almahara.2025.0111-09

Nasr, I. (2026). Balancing formative and summative assessment in IB MYP Arabic B: A qualitative case study of teachers' practices and perspectives. *An-Najah University Journal for Research – B (Humanities)*. https://doi.org/10.35552/0247.a2830

Rakhlin, N. V., Aljughaiman, A., & Grigorenko, E. L. (2021). Assessing language development in Arabic: The Arabic language: Evaluation of Function (ALEF). *Applied Neuropsychology: Child*, *10*(1), 37–52. https://doi.org/10.1080/21622965.2019.1596113

Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. (2024). Item analysis: The impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, *24*(1), 1–7. https://doi.org/10.1186/s12909-024-05433-y

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Schmucker, R., & Moore, S. (2026). The impact of item-writing flaws on difficulty and discrimination in item response theory. *Computers and Education: Artificial Intelligence*, *11*, 100632. https://doi.org/10.1016/j.caeai.2026.100632

Sharma, L. R. (2021). Analysis of difficulty index, discrimination index and distractor efficiency of multiple-choice questions of speech sounds of English. *International Research Journal of MMC*, *2*(1), 15–28. https://doi.org/10.3126/irjmmc.v2i1.35126

Sireci, S. G., & Rodriguez, G. (2022). *Validity in educational testing* (pp. 1–9). https://doi.org/10.4324/9781138609877-ree180-1

Subando, J. (2025). A multidimensional item response theory approach in the item analysis of Arabic language tests in Madrasah Aliyah. *Jurnal Penelitian dan Evaluasi Pendidikan*, *29*(2), 271–284. https://doi.org/10.21831/pep.v29i2.90877

Yu, G., & Xu, J. (Eds.). (2024). *Language test validation in a digital age*. Cambridge University Press.

Zeinoun, P., Iliescu, D., & El Hakim, R. (2022). Psychological tests in Arabic: A review of methodological practices and recommendations for future use. *Neuropsychology Review*, *32*(1), 1–19. https://doi.org/10.1007/s11065-021-09476-6

Zhao, H., & Aryadoust, V. (2025). A meta-analysis of the reliability of second language reading comprehension assessment tools. *Studies in Second Language Acquisition*, *47*(1), 388–416. https://doi.org/10.1017/S0272263124000627