



Toward a Standardized Arabic Proficiency Test to Support Graduate Academic and Communicative Skill in Indonesian Higher Education

نحو اختبار معياري لإتقان اللغة العربية لدعم المهارات الأكاديمية والتواصلية لطلاب الدراسات العليا في التعليم العالي الإندونيسي

Hikmah Maulani^{1*}, Abdul Kodir², Asep Supianudin³, Nanang Kosim⁴, Ashinida Aladdin⁵

¹Universitas Pendidikan Indonesia, Bandung, Indonesia

^{2,3,4}Universitas Islam Negeri Sunan Gunung Djati, Bandung, Indonesia

⁵Universiti Kebangsaan Malaysia, Malaysia

*Corresponding E-mail: hikmahmaulani@upi.edu

ABSTRACT

This study aims to analyze the gap between theoretical constructs and empirical practices in Arabic language proficiency tests in Indonesian higher education, as well as to identify the direction of standardization that is able to bridge the orientation of Arabic language proficiency in Higher Education in two categories, namely Arabic Language Education and Arabic Literature. This research uses a qualitative approach with an a priori-oriented case study method, conducted at two public general universities in Indonesia, one specializing in Arabic Language Education and the other in Arabic Literature. Data were collected through document analysis (covering two curricula, test instruments, syllabi, and learning outcomes) and semi-structured interviews with two lecturers who also serve as test developers. The analysis was conducted using theory-driven thematic analysis to examine the constructs of proficiency, skill representation, and standardization practices. The findings indicate significant gaps in three main aspects: (1) constructs that are still oriented towards language knowledge rather than language use, (2) an imbalance in skill representation with a dominance of qira'ah and a lack of productive skills, and (3) the absence of calibrated standards across institutions. Furthermore, there is a dichotomy between the communicative (Arabic Language Education) and analytical (Arabic Literature) approaches that have not been integrated into the test design. This research offers a conceptual contribution by positioning Arabic language proficiency as an integrative competency that combines communicative and analytical dimensions, while simultaneously bridging two previously separate disciplines. This study provides a basis for the development of more representative, contextual, and standardized proficiency tests at the higher education level.

Keywords: Academic Standards; academic skill; assessment; policy in higher education.

مستخلص البحث

تهدف هذه الدراسة إلى تحليل الفجوة بين البنى النظرية والممارسات التطبيقية في اختبارات كفاءة اللغة العربية بمؤسسات التعليم العالي الإندونيسي، فضلاً عن تحديد اتجاه المعايير القادرة على تجسير توجه كفاءة اللغة العربية في التعليم العالي ضمن فئتين، هما: تعليم اللغة العربية، وأدب اللغة العربية. يستخدم هذا البحث مخرجاً نوعياً (كيفياً) باعتماد منهج دراسة الحالة الموجهة مسبقاً (a priori)، وقد أُجري في جامعتين حكوميتين عامتين في إندونيسيا؛ إحداهما متخصصة في تعليم اللغة العربية والأخرى في الأدب العربي. وُجمعت البيانات من خلال تحليل الوثائق (التي شملت مناهج دراسيين، وأدوات الاختبار، والتوصيفات الدراسية، ومخرجات التعلم)، بالإضافة إلى المقابلات شبه المنظمة مع اثنين من المحاضرين من مطوري الاختبارات. وتُشير النتائج إلى وجود فجوات كبيرة في ثلاثة جوانب رئيسية تشمل البنى التي لا تزال مُوجَّهة نحو المعرفة اللغوية بدلاً من استخدام اللغة، والخلل في تمثيل المهارات مع هيمنة مهارة القراءة وضعف المهارات الإنتاجية، بالإضافة إلى غياب المعايير المُقنَّنة بين المؤسسات؛ علاوة على ذلك، ثمة انقسام بين المقاربتين التواصلية والتحليلية اللتين لم يتم دمجهما في تصميم الاختبار. وتُقدِّم هذه الدراسة مساهمة مفاهيمية من خلال وضع كفاءة اللغة العربية ككفاية تكاملية تجمع بين البُعدين التواصلية والتحليلية لتُجسِّر بذلك بين تخصصين كانا

Citation:

Maulani, H., Kodir, A., Supianudin, A., Kosim, N., & Aladdin, A. (2026). Toward a Standardized Arabic Proficiency Test to Support Graduate Academic and Communicative Skill in Indonesian Higher Education. *Al-Muyassar: Journal of Arabic Education*, 5 (3): 608–628

منفصلين، مما يُوفّر أساساً لتطوير اختبارات كفاءة أكثر تمثيلاً، وسياًفاً، ومعياريّةً على مستوى التعليم العالي.
الكلمات المفتاحية: المعايير الأكاديمية، المهارات الأكاديمية، التقويم، السياسة التعليمية في التعليم العالي.

INTRODUCTION

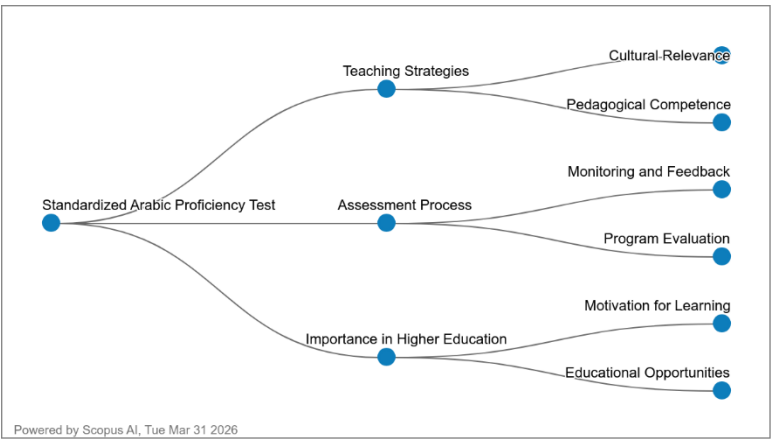
Language testing practices in higher education are never neutral (Fulcher, 2013; Cook, 2017; Carlsen & Rocca, 2022). They reflect academic policies (Cumming, 2009), curriculum ideologies (Reay, 2018), social (Roever & McNamara, 2006), and epistemological constructions of what constitutes “competence” (Fulcher, 2013). In the context of Arabic language teaching in Indonesian public universities, proficiency tests have become a strategic instrument for measuring graduate achievement and establishing minimum student competency standards (Gazali & Saefuloh, 2023; Maulani et al., 2024). Arabic language proficiency has become a crucial competency for students at various universities in Indonesia (Kamal, 2025).

In many Indonesian universities, Arabic language proficiency tests lack a uniform standard to effectively evaluate student competencies (Wahyu Pratama et al., 2024), despite the strategic role testing plays in shaping curriculum policy and graduate outcomes (Oliver, 2013). While local institutions often adapt the Common European Framework of Reference for Languages (CEFR) (Musthofa, 2022; Mohamed, 2023; Miqdarul Khoir Syarofit et al., 2025) and global frameworks vary from ACTFL standards to ALPT and the Arab Academy (Liskin-Gasparro, 1984) a universally accepted benchmark for non-native contexts remains unestablished. Unlike other foreign languages that possess well-defined content standards (Yonkers, 1996), this absence of standardized testing hinders Arabic language development. In Indonesia, this gap is particularly critical as Arabic has evolved beyond a general subject into distinct undergraduate, master's, and doctoral programs divided into two primary sub-disciplines: Arabic Language Education and Arabic Literature (R. Taufiqurrochman & Lubna Farah, 2024). Consequently, establishing standardized proficiency tests tailored to this disciplinary dichotomy is vital to accurately assess and support graduates' academic and communicative skills.

Thus, developing a standardized Arabic language proficiency test in higher education in Indonesia is a complex but necessary endeavor (see Figure 1), given the challenges and opportunities in Arabic language education across the country. While no abstracts directly address the creation of such a test, several insights from the studies

presented can serve as a reference in its development. For example, a 2020 study developed an online test to facilitate the pandemic (Fauzi et al., 2020). In 2021, Rakhlin et al. identified early language difficulties in the spoken and written domains, and from this, developed a series of assessment tests called the Arabic Language: Functional Evaluation (ALEF), intended for children aged 3 to 11. In 2023, a test was developed specifically for Arabic semantics as a valid and reliable test applicable to assess semantic development in Arabic-speaking preschool children in non-speaking environments (Abd-Elmoneim et al., 2023).

Question construction analyses have also been extensively conducted (Baso et al., 2023; Ghani & Daud, 2023; Maulani et al., 2024; Koto et al., 2024). The focus of language model evaluation has begun to shift toward reasoning and knowledge-intensive tasks, driven by advances in the initial training of large Modern Standard Arabic (MSA)-based models. Furthermore, developments in academic policy for Arabic language tests are increasingly being highlighted and studied more specifically (Abdelbaset & Jabari, 2020; Al-Bataineh, 2021; Alkhateeb & Bouherar, 2023)



Source: Scopus AI

Figure 1. Arabic language proficiency test research map

The development of a standardized Arabic proficiency test in Indonesian higher education must directly address the fundamental dichotomy between Arabic Language Education (oriented toward communicative competence) and Arabic Literature (oriented toward analytical and textual mastery). Currently, this rift poses a major challenge: a single, generic test risk either under-assessing the pedagogical and oral fluency needed in education programs or failing to measure the complex literary analysis and philological skills required in literature tracks. To overcome this, tests must integrate international frameworks like the CEFR with localized adaptations that account for

Arabic diglossia, institutional nuances, and regional disparities. Updating curricula, elevating teacher competencies, and leveraging digital platforms alongside dynamic assessment models such as portfolio-based evaluations are critical steps toward developing an instrument that is both valid and tailored to these distinct academic trajectories.

This research carefully answers the question, namely

Q1 How are Arabic language proficiency tests designed in higher education institutions in Indonesia?

Q2 To what extent do the tests reflect standardized language proficiency?

Therefore, this research has an important practical role in breaking the deadlock in the direction of development of Arabic language tests for non-speakers in Indonesian universities. To strengthen the introduction, the following are the theoretical frameworks used in this study.

Language Testing Theory

A language proficiency test is an instrument designed to measure a person's level of ability to use a language in general, regardless of a specific curriculum or learning program (Bachman & Clark, 1987). In language testing studies, proficiency tests differ from achievement tests. Achievement tests focus on material learned in a program, while proficiency tests aim to measure language ability more broadly and contextually (Abedi, 2002).

These developments indicate that the practice of language evaluation continues to expand, not only in the scope of use, but also in the complexity of the approaches used in designing assessment instruments (Chalhoub-Deville, 1997). Although the terms competency and proficiency are used, there is no clear consensus regarding their definitions; Therefore, competence is understood within Noam Chomsky's framework, while communicative language ability includes language proficiency and communicative proficiency in language use (Llurda, 2000),.

However, a literature review shows that the L2 proficiency models used by test developers have significant component diversity (Brown, 1995). Each model offers a different perspective on what is meant by language proficiency, whether in terms of linguistic competence, communicative ability, or strategic aspects of language use (McNamara et al., 1996). This diversity ultimately shows that there is no complete agreement among experts on how language proficiency should be represented in a test

(Bernard Spolky, 1992).

Arabic as a Foreign Language Assessment

Arabic is spoken as a mother tongue by over 300 million people across two continents: Africa and Asia. Furthermore, it is an official language in 26 countries worldwide. Given its large number of native speakers and the large number of countries where it is spoken as a mother tongue, Arabic was chosen as one of the six official languages of the United Nations (Gebril & Taha-Thomure, 2013). Arabic is a Semitic language with 28 letters and is written from right to left. Modern Standard Arabic (MSA) is used in the media and various formal public contexts and is generally understandable to Arabs who have completed basic education (Alzahrani, 2024).

A review of various Arabic language tests has revealed a wide variety of assessments used for various purposes in academic settings such as admissions, placement, and exit decisions. In non-academic contexts, these tests are primarily used for recruitment or promotion purposes, particularly in Arabic Foreign Language (AFL) settings. Recent trends in language assessment, which embrace new assessment functions beyond screening, should be considered in the development of Arabic language tests (Gebril & Taha-Thomure, 2013).

However, without proper evaluation procedures, it is difficult to empirically verify the value of language theories. Therefore, testing should be viewed as a means to an end: improved teaching practices, better teaching materials, and fair evaluation of learners. Language teachers are always involved in testing in one way or another. Learning and its measurement are closely intertwined. Without a sufficient understanding of the various approaches to language teaching and the differences between them, measurement procedures can be haphazard and unfocused (Mohamed, 2023). Consequently, Arabic language learners face challenges due to the limited internationally accepted framework of competency standards, which limits the validity and comparability of tests (Rakhlin et al., 2021).

Language Proficiency Standards Framework

Proficiency is a complex and useful concept for measuring language learning outcomes, tracking student progress, and providing a common language for students, instructors, administrators, and other stakeholders to discuss these aspects (Harsch,

2018). Standard-setting practices are often used to link test scores to descriptions of language ability, as this is one way to assign meaning to those scores. However, there are numerous standard-setting methods and procedures, and while there are suggested guidelines for how to apply them, not all studies adhere to them consistently or to the same degree (Fleckenstein et al., 2020).

In many areas of industry, society, and research, standards are used to simplify communication and standardize measurement. In linguistics, standards are also used for this purpose (Deygers & Van Gorp, 2015). For example, various states are currently working to develop or adopt new English Language Proficiency (ELP) assessments aligned with new academic standards to meet accountability requirements. In this transition, it is crucial to identify areas for improvement in current ELP assessment practices and discuss ways to implement effective ELP assessments to better serve the needs of English Language Learner (ELL) students (Wolff & Costa, 2017).

RESEARCH METHOD

This research employed a qualitative case study approach (Duff, 2007) guided by an *a priori* perspective (Casullo, 2009), a strategy where the analytical process is structured around a pre-formulated conceptual framework. The case study design was selected to facilitate an in-depth exploration of the construction and standardization practices of Arabic proficiency tests within the authentic context of Indonesian higher education, particularly when the boundaries between the phenomenon and its context are intertwined (Glenn A. Bowen, 2009).

Operationally, this conceptual framework integrates the construct of language proficiency, the balance of the four macro-skills, and core principles of assessment design, including representation and standardization. By grounding the analysis in these pre-established dimensions, the approach ensures a systematic evaluation of test structures while remaining responsive to dynamic empirical findings in the field (Tomaszewski et al., 2020).

t in the Arabic Language Education and Arabic Literature study programs at universities. Case selection was conducted purposively to obtain a rich and informative variety of assessment practices (information-rich cases) (Malterud et al., 2016).

Data collection was conducted using several techniques: (1) document analysis of test instruments, syllabi, and graduate learning outcomes; and (2) semi-structured

interviews with test developers. This approach enabled researchers to understand both the conceptual aspects and implementation practices in test construction. To enhance data validity, triangulation of sources and methods was used (Denzin, 2009).

Data analysis was conducted using a priori category-based thematic analysis, in which data were coded and classified based on predetermined constructs, such as communicative competence, integration of vocabulary and grammar, and balance of language skills. This approach aligns with theory-driven thematic analysis, which allows for systematic exploration of data based on a conceptual framework (Thelwall & Nevill, 2021).

To ensure the credibility and validity of the findings, this study applied the principle of trustworthiness through: (1) data triangulation, (2) expert validation, and (3) an iterative and reflective analysis process. This approach aligns with the criteria of credibility, dependability, and confirmability in qualitative research (Verner & Abdullah, 2012).

With this approach, the study focused on conceptual analysis and the representational practices and standardization of proficiency tests, without involving statistical analysis of the items. This allowed for a more in-depth exploration of how the proficiency construct is understood, represented, and implemented in the real context of higher education.

From this perspective, statistical validity testing will not adequately address the issue if the underlying construct of the test is not precisely defined. Therefore, this study prioritized construct analysis (construct validity at the conceptual level) by examining the relationship between curriculum objectives, teaching materials, and evaluation formats. This approach enabled a more in-depth understanding of how reading ability is conceptualized, transformed, and represented in testing practices in higher education.

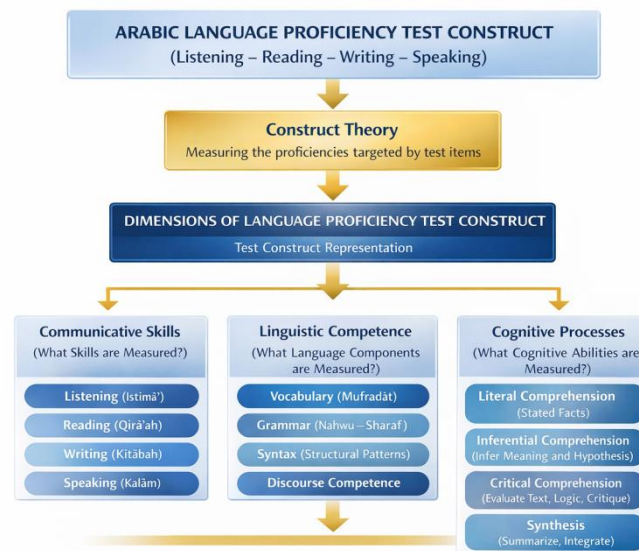


Figure 2. conceptual framework (Source: author's own work)

This framework integrates the principles of language testing theory with the construct of Arabic language proficiency to develop a standardized proficiency test model. The construct of language proficiency consists of four core skills: *istimā'*, *qirā'ah*, *kitabah*, and *kalam*, supported by linguistic competencies in vocabulary, grammar, and discourse.

These constructs are then translated into test design through specification of question formats, test blueprints, and levels of cognitive demand. The results of language proficiency measurements are then mapped to a proficiency scale aligned with international frameworks such as the Common European Framework of Reference for Languages, which divides language proficiency into six levels from A1 to C2. This framework is expected to serve as the basis for developing a standardized Arabic language proficiency test for the Indonesian context, thereby improving the consistency and validity of Arabic language proficiency assessment in higher education.

RESULT AND DISCUSSION/ نتائج البحث والمناقشة

What a Proficiency Test Should be?

A key point to emphasize is the distinction between achievement tests and diagnostic tests. In the context of higher education, a test is administered to measure students' Arabic language skills, both as learning outcomes and as graduate outcomes (Al-Qatawneh et al., 2021), not simply as a measure of course learning outcomes. Rather, it serves as a benchmark for graduate proficiency (Wang et al., 2012).

Proficiency tests are evaluation instruments designed to measure general and

comprehensive language skills (North, 1995), without being tied to a specific learning program or curriculum. Unlike learning outcome tests, which focus on mastery of previously taught material, proficiency tests aim to reflect a person's actual level of language use in various contexts (Bloom, 2008). Therefore, these tests must be able to represent language skills authentically and functionally.

In their development, proficiency tests must be based on a clear and well-defined construct of language ability. This construct encompasses not only linguistic aspects but also communicative, interpretive, and strategic language use. Thus, the test measures not only language knowledge but also the ability to use language to understand and convey meaning in real-life situations.

Furthermore, a good proficiency test must meet the main principles of language assessment, such as validity, reliability, and authenticity (Bachman, 2001). Validity ensures that a test accurately measures the intended ability (Reay, 2018), while reliability ensures the consistency of measurement results (Wesolowski & Wind, 2019). Authenticity requires that test tasks reflect real-life language use (Wu et al., 2017), so that test results have meaning relevant to communication needs.

Proficiency tests must also be able to represent language ability in the form of tasks or items that align with the construct being measured (Abdelbaset & Jabari, 2020). In the context of language skills, for example, tests need to measure various levels of understanding, from literal comprehension to interpretive and evaluative comprehension (Basaraba et al., 2013). This suggests that test design must consider the relationship between the ability construct and its representation in the questions.

Furthermore, proficiency tests ideally have a clear leveling system to indicate language ability development and can be used across contexts and institutions, such as the standards established by the CEFR and ACTFL. With these characteristics, proficiency tests not only function as measuring tools, but also as indicators of ability that can be used for various academic and professional purposes.

Thus, a good proficiency test is a test that is able to measure language abilities as a whole, is based on clear constructs, and is relevant to language use in real contexts. These principles are an important basis in evaluating and analyzing the design of Arabic language proficiency tests, especially in reading skills, which is the focus of this research.

Emprical Description

Indonesia, with its two study programs: Arabic language education and Arabic literature, should have a comprehensive framework to arbitrate the quality of language skills for Arabic language graduates studying in higher education (see Table 1).

Table 1. Competency achievement framework for two Arabic language study programs in Indonesia

Study Program	Orientation Competency	Description
Arabic education	Pedagogical-didactic (teaching, learning design)	Understanding learning discourse, classroom instruction, and educational dialogue
	<i>Istima'</i> (listening)	
	<i>Kalam</i> (speaking)	
	<i>Qirā'ah</i> (reading)	
Arabic Literature	<i>Kitābah</i> (writing)	
	Analytical-critical (study of texts, literature, and culture)	Understanding narratives and descriptions in the general context of Arabic culture
	<i>Istima'</i> (listening)	
	<i>Kalam</i> (speaking)	
	<i>Qirā'ah</i> (reading)	
	<i>Kitābah</i> (writing)	

Source: author's own work from curriculum documents of the Indonesian University of Education and Hasanuddin University

Considering the principles of proficiency testing, this division into two programs does not limit the essence of a proficiency standard; rather, this standard further serves as a qualification for program graduates, whether in education or literature, to possess the same proficiency.

In Indonesian language learning practices, another aspect of language proficiency is the tendency toward nahwu-sharaf (literary grammar), or the equivalent of syntax and morphology, and vocabulary memorization. Mastery of nahwu-sharaf and mufradat constitutes a core linguistic competency that is a prerequisite for the development of the four language skills. Nahwu-sharaf functions as a system for organizing form and structure, while mufradat serves as a carrier of meaning. Both interact simultaneously in receptive (*istimā'*, *qirā'ah*) and productive (*kalām*, *kitabah*) processes, so the quality of

their mastery directly determines the overall level of language proficiency (see Figure 3).

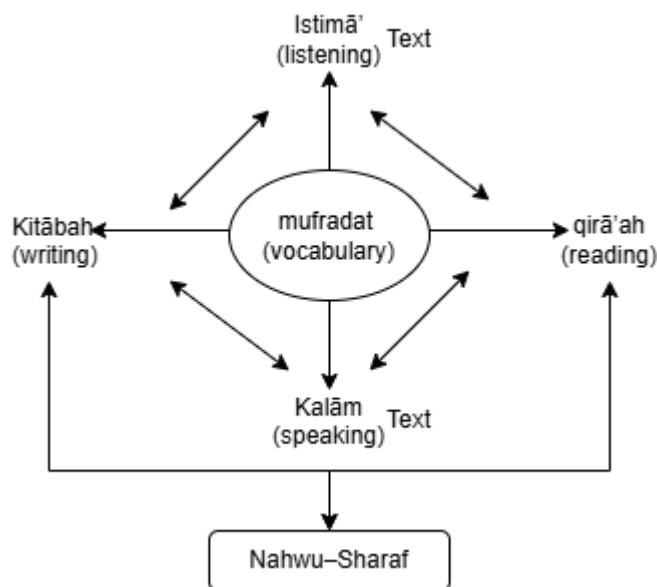


Figure 3. Interactive relationship of Arabic language elements applied in the field (course) (source: author's own work)

Vocabulary is positioned as the center of linguistic competence, serving as the primary source for constructing meaning. This position demonstrates that vocabulary mastery is fundamental to the development of overall language skills. From vocabulary, there is a two-way relationship to the four language skills. Istimā' (listening) and qirā'ah (reading) represent receptive skills that enable learners to receive and understand language input, while kalam (speaking) and kitabah (writing) are productive skills that require the ability to express meaning orally and in writing. This two-way relationship emphasizes that vocabulary mastery is not only a prerequisite for language skills but also continues to develop through the practice of using language in various contexts.

Meanwhile, nahwu-sharaf (literary grammar) is positioned as a structural foundation that primarily supports productive skills. Its existence demonstrates that speaking and writing skills depend not only on vocabulary mastery but also on the ability to organize words into grammatically correct structures. The flow connecting kalam and nahwu-sharaf indicates that the practice of actively using language contributes to strengthening the understanding of grammatical structures. Thus, this model emphasizes that Arabic language learning takes place in an integrative manner, where vocabulary, structure, and language skills interact and develop simultaneously.

How Design Happens in Practice

In practice at Indonesian universities, the design of Arabic language proficiency tests is often not fully based on a systematic conceptual framework. For example, there

are no standards from Arab countries for Arabic language proficiency criteria (Alzahrani, 2024). Rather than starting from a clearly defined ability construct, tests tend to be pragmatic, following established practices or referring to material taught in class. As a result, the resulting tests reflect more on the content of the learning process than the language skills they are intended to measure. The following is an overview of tests that are legally administered to university students in Indonesia.

Table 2. Examples of Arabic language proficiency tests from Indonesian institutions

Test Name			Scoring-leveling	Interpretation	Program Participant Coverage	Program Type
TOAFL	UIN	Syarif Hidayatullah	Minimum score of 500	is required for graduation	Language program standards tend to be higher than general program standards	Online
<i>Test intended for university students</i>						
IKHFA	Test	Universitas Pendidikan Indonesia	Minimum score of 500 is required for bachelor's degree programs, 550 is required for master's degree programs.	is required for graduation	College students	Online
<i>Test intended for university students</i>						
TOAFL	IBT	Al Azhar Center Pare	Scores range from 210 to 680.	Grades range from A for excellent and D for the	General	Online
<i>Test intended specifically for non-Arabic speakers</i>						

Test Name	Scoring- leveling	Interpretation	Program Participant Coverage	Program Type
worst				

Furthermore, the absence of explicit test specification documents indicates that the design process often occurs implicitly and undocumented. Item developers generally rely on intuition, teaching experience, or previous sample items, without going through a structured construct derivation process. This condition results in a weak link between learning outcomes, teaching materials, and assessment methods, thus questioning construct validity.

In the context of testing, dominant practice still strongly favors testing structural aspects of language, such as isolated vocabulary comprehension or grammatical analysis, rather than measuring the ability to comprehend complete meaning. The items often require less interpretive or evaluative skills than simply recognition of language forms. This indicates that reading ability is reduced to a decoding activity, rather than a process of constructing meaning.

Empirical descriptions of Arabic language proficiency standards in the context of higher education in Indonesia indicate a tendency for a disconnect between global theoretical frameworks and local practices. Although some institutions have adopted reference frameworks such as the Common European Framework of Reference (CEFR) (Nurdianto et al., 2025), their implementation remains partial and not fully calibrated to the characteristics of Arabic learners as a foreign language.

Based on field findings, proficiency standards are often implicitly determined through coursework, rather than through operationally defined language proficiency constructs. As a result, students' proficiency levels (e.g., equivalent to B1 or B2) are not always supported by measurable performance indicators, particularly in productive skills such as kalam and kitabah. This indicates a gap between proficiency level labels and students' actual competencies.

Empirically, students' Arabic language mastery tends to be unbalanced across proficiencies. Qira'ah skills are relatively more developed than istimā' and kalam, which are often limited to the classroom context and less supported by authentic input. Furthermore, the dominance of a structural approach in learning results in theoretically

strong mastery of nahwu and sharaf, but not yet fully integrated into communicative language use. Furthermore, limited vocabulary is a major factor hindering text comprehension and fluent language production.

Other findings indicate that the construction of proficiency tests used in some study programs does not fully meet the principles of construct validity. The instruments used tend to measure language knowledge rather than language use. For example, multiple-choice questions that emphasize grammatical rules are more dominant than performative tasks that reflect language use in real-world contexts.

Furthermore, standardization and calibration processes between institutions are still limited. The absence of a specific national benchmark for Arabic results in significant variation between universities, both in terms of achievement targets and measurement methods. This impacts the low comparability of proficiency test results between institutions.

Thus, empirically, it can be concluded that Arabic proficiency standards in Indonesia are still in the development stage, with the main challenges being: (1) clarity of constructs and level indicators, (2) balance between language skills, (3) integration of nahwu-sharaf and mufradat in language use, and (4) development of valid, performance-based assessment instruments. Therefore, systematic efforts are needed to develop proficiency standards that are contextual, measurable, and aligned with learning practices in the field.

Representation Analysis

Representation analysis in the development of Arabic language proficiency tests focuses on the extent to which the language ability constructs tested truly reflect holistic and contextual language competencies. In the context of higher education in Indonesia, this representation is crucial because proficiency tests serve not only as measuring tools but also as a reflection of the orientation of Arabic language learning itself.

Empirically, the representation of Arabic language ability in existing instruments tends to be partial. Recitation skills often receive a more dominant portion than other skills, primarily due to the ease of constructing text-based items. Conversely, productive skills such as kalam and kitabah are underrepresented proportionally, both in terms of quantity and task complexity. This results in the resulting tests representing language knowledge rather than language use.

Furthermore, the linguistic aspects represented in the tests also show an

imbalance. Mastery of grammar and grammar is often measured separately through decontextualized rule questions, while vocabulary is tested in the form of lexical meaning recognition without integration with discourse usage. As a result, tests do not fully reflect communicative competence, which requires the integration of form, meaning, and use (Lichtman & VanPatten, 2021).

In terms of context, the representation of tasks in tests is also still limited to narrow academic situations and does not adequately reflect the variety of language use in real life. The texts and tasks used tend to be inauthentic and less relevant to students' communicative needs in the global era, resulting in low ecological validity.

Efforts to standardize Arabic language proficiency tests in Indonesia require a paradigm shift in construct representation. Tests need to be designed to represent language ability holistically by considering: (1) balance among the four skills, (2) integration between *mufradat* and *nahwu-sharaf* in the context of language use, and (3) the use of authentic tasks that reflect real-life communication situations.

Thus, this representation analysis confirms that the primary challenge in developing an Arabic language proficiency test lies not only in the technical aspects of instrument construction, but also in how the construct of language ability is represented in a valid, comprehensive, and contextual manner. This is a fundamental step toward realizing a standardized and comparable proficiency test nationally and internationally.

Standardization and Alignment

Standardization and harmonization in the development of Arabic language proficiency tests in Indonesian higher education still face several fundamental challenges. Empirically, existing assessment practices demonstrate significant variation across institutions, both in terms of defining the ability construct, establishing proficiency levels, and the format and quality of the instruments used. This condition results in poor comparability of test results and weakens the test's function as a valid and accountable measurement tool.

One key issue lies in the lack of agreement on the definition of the Arabic language proficiency construct. In many cases, language ability is still represented partially, with a greater emphasis on aspects of linguistic knowledge such as grammar and grammar, while the ability to use language in communicative contexts has not been adequately accommodated. As a result, the tests developed often do not reflect complete language proficiency.

In terms of proficiency levels, there is a tendency to use labels such as beginner, intermediate, and advanced without being supported by operational and measurable ability descriptors. Some institutions have begun to refer to international frameworks such as the CEFR, but their implementation is inconsistent and often not accompanied by adequate adaptation to the characteristics of the Arabic language and the context of learners in Indonesia. This creates a mismatch between the stated levels and the actual abilities of participants.

Furthermore, the aspect of standardization is also closely related to the quality of test instruments. Many tests are still dominated by multiple-choice questions that focus on grammatical rules, thus measuring language knowledge rather than language use. Productive skills such as speaking and writing tend to be underrepresented, both due to technical limitations and the lack of clear and calibrated assessment standards.

In the effort to achieve standardization, strategic steps are needed, including: (1) formulating a comprehensive proficiency construct that encompasses the language use dimension, (2) developing clear, measurable, and contextual level descriptors, (3) improving instrument quality through a performance-based approach, and (4) implementing empirical analysis procedures to ensure test validity and reliability.

Furthermore, harmonization between institutions is equally important. Without coordination and benchmarking mechanisms, developed standards will remain local and difficult to compare. Therefore, collaboration between universities, language centers, and other stakeholders is necessary to build a shared understanding of Arabic language proficiency standards that can be widely applied.

Therefore, standardization and alignment in this context are gradual processes that require the integration of conceptual clarity, empirical consistency, and institutional agreement. This effort serves as a crucial foundation for developing a more standardized, credible, and relevant Arabic language proficiency testing system for higher education in Indonesia. The pursuit of a standardized Arabic language proficiency test is not only a technical measurement issue but also a paradigm shift in understanding and representing language proficiency. This research provides a conceptual and empirical foundation for developing a more integrative, credible, and relevant assessment system for higher education in Indonesia.

CONCLUSION

This study confirms that the development of Arabic language proficiency tests in Indonesian higher education still faces fundamental challenges regarding construct representation, skill balance, and standardization. Conceptually, a persistent gap remains between the theoretical understanding of proficiency as a communicative ability and actual assessment practices, which remain heavily oriented toward discrete language knowledge. Furthermore, this study highlights the structural dichotomy between the communicative orientation of Arabic Language Education and the analytical orientation of Arabic Literature, both of which have yet to be integrated into a unified proficiency framework.

Practically and policy-wise, these findings underscore the urgent need for curriculum developers and institutional test boards to redesign proficiency models that bridge this academic divide. Standardized testing must accommodate both communicative fluency and analytical mastery to ensure valid assessment across both degree tracks.

While this study provides in-depth qualitative insights from two public universities, its empirical scope is limited to specific institutional contexts and document sets. To advance this research, specific next steps should focus on:

First, developing a unified framework matrix: designing a prototype test blueprint that operationalizes both communicative and literary-analytical competencies into distinct proficiency bands.

Second, large-scale validation studies: conducting multi-site quantitative validation tests across a broader range of Islamic and public universities in Indonesia to establish psychometric reliability and national benchmarks.

ACKNOWLEDGEMENT

This research is funded by the Universitas Pendidikan Indonesia through the research scheme/program of Riset Afiriasi Dosen with contract No 179/UN40.F7/PT.01.01/2026.

REFERENCES

- Abd-Elmoneim, D. M., Ghandour, H. H., Elrefaie, D. A., & Khodeir, M. S. (2023). Development of an Arabic test for assessment of semantics for the Arabic-speaking children: the Arabic semantic test. *Egyptian Journal of Otolaryngology*, 39(1). <https://doi.org/10.1186/s43163-023-00405-3>
- Abdelbaset, T. E. M., & Jabari, N. A. . (2020). Arabic language teaching evaluation process for non-native speakers. *International Journal of Education and Learning*, 2(2), 105–110. <https://doi.org/10.31763/ijele.v3i1.157>
- Abedi, J. (2002). Standardized achievement tests and english language learners: Psychometrics issues. *International Journal of Phytoremediation*, 21(1), 231–257. https://doi.org/10.1207/S15326977EA0803_02
- Al-Bataineh, A. (2021). Language policy in higher education in the United Arab Emirates: proficiency, choices and the future of Arabic. *Language Policy*, 20(2), 215–236. <https://doi.org/10.1007/s10993-020-09548-y>
- Al-Qatawneh, S. S., Alsalhi, N. R., Eltahir, M. E., & Siddig, O. A. (2021). The representation of multiple intelligences in an intermediate Arabic-language textbook, and teachers' awareness of them in Jordanian schools. *Heliyon*, 7(5), e07004. <https://doi.org/10.1016/j.heliyon.2021.e07004>
- Alkhateeb, H., & Bouherar, S. (2023). Sustainability and educational language policy in Arab higher education: findings from Q research. *Current Issues in Language Planning*, 24(5), 573–595. <https://doi.org/10.1080/14664208.2022.2155925>
- Alzahrani, A. (2024). LexArabic: A receptive vocabulary size test to estimate Arabic proficiency. *Behavior Research Methods*, 56(6), 5529–5556. <https://doi.org/10.3758/s13428-023-02286-z>
- Bachman, L. F. (2001). Fundamental considerations in language testing / Lyle F. Bachman. In *Fundamental considerations in language testing* (pp. 81–109). <https://openurl.ebsco.com/contentitem/cat00040a:bilk.2013742?sid=ebsco:plink:crawler&id=ebsco:cat00040a:bilk.2013742&crl=c>
- Bachman, L. F., & Clark, J. L. D. (1987). The Measurement of Foreign/Second Language Proficiency. *The ANNALS of the American Academy of Political and Social Science*, 490(1), 20–33. <http://hjb.sagepub.com.proxy.lib.umich.edu/content/9/2/183.full.pdf+html>
- Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing*, 26(3), 349–379. <https://doi.org/10.1007/s11145-012-9372-9>
- Baso, Y. S., Murtadho, N., -, S., Maulani, H., Agussalim, A., -, H., Fadlan, A., & Ramadhan, I. (2023). Reducing Cheating in Online Exams Through the Proctor Test Model. *International Journal of Advanced Computer Science and Applications*, 14(1), 366–371. <https://doi.org/10.14569/ijacsa.2023.0140139>
- Bernard Spolky. (1992). Communicative competence and language proficiency testing. *Asian Studies Review*, 15(3), 30–37. <https://doi.org/10.1080/10357823.1992.9755377>
- Bloom, M. (2008). Second language composition in independent settings: Supporting the writing process with cognitive strategies. In *Language Learning Strategies in Independent Settings*. <https://doi.org/10.21832/9781847690999-008>
- Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12(1), 1–15. <https://doi.org/10.1177/026553229501200101>
- Carlsen, C. H., & Rocca, L. (2022). Language test activism. *Language Policy*, 21(4), 597–616.

- <https://doi.org/10.1007/s10993-022-09614-7>
- Casullo, A. (2009). Analyzing a priori knowledge. *Philosophical Studies*, 142(1), 77–90. <https://doi.org/10.1007/s11098-008-9302-5>
- Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing*, 14(1), 3–22. <https://doi.org/10.1177/026553229701400102>
- Cook, W. R. A. (2017). More vision than renaissance: Arabic as a language of science in the UAE. *Language Policy*, 16(4), 385–406. <https://doi.org/10.1007/s10993-016-9413-3>
- Cumming, A. (2009). Language assessment in education: Tests, curricula, and teaching. *Annual Review of Applied Linguistics*, 29, 90–100. <https://doi.org/10.1017/S0267190509090084>
- Denzin, N. K. (2009). The elephant in the living room: Or extending the conversation about the politics of evidence. *Qualitative Research*, 9(2), 139–160. <https://doi.org/10.1177/1468794108098034>
- Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, 32(4), 521–541. <https://doi.org/10.1177/0265532215575626>
- Duff, P. (2007). *Case Study Research in Applied Linguistics* (1st Editio). Routledge.
- Fauzi, M. F., Murdiono, M., Anindiati, I., Nada, A. L. I., Khakim, R. R., Mauludiyah, L., & Thoifah, I. (2020). Developing Arabic Language Instructional Content in Canvas LMS for the Era and Post Covid-19 Pandemic. *Izdihar: Journal of Arabic Language Teaching, Linguistics, and Literature*, 3(3), 161–180. <https://doi.org/10.22219/jiz.v3i3.15017>
- Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R. J., & Köller, O. (2020). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing*, 43(July 2018), 100420. <https://doi.org/10.1016/j.asw.2019.100420>
- Fulcher, G. (2013). Philosophy and Language Testing. *The Companion to Language Assessment*. <https://doi.org/10.1002/9781118411360.wbcla032>
- Gazali, E., & Saefuloh, H. (2023). Development of an Arabic Receptive Proficiency Test Instrument Based on the Common European Framework of Reference for Languages. *Al-Ta'rib: Jurnal Ilmiah Program Studi Pendidikan Bahasa Arab IAIN Palangka Raya*, 11(2), 293–308. <https://doi.org/10.23971/altarib.v11i2.6721>
- Gebril, A., & Taha-Thomure, H. (2013). Assessing ARABIC. *The Companion to Language Assessment*. <https://doi.org/10.1002/9781118411360.wbcla065>
- Ghani, M. T. A., & Daud, W. A. A. W. (2023). The Impact of Digital Game-Based Learning Towards Arabic Language Communication. *Jurnal Komunikasi: Malaysian Journal of Communication*, 39(1), 407–424. <https://doi.org/10.17576/JKMJC-2023-3901-23>
- Glenn A. Bowen. (2009). Document analysis as a qualitative research method. In *Qualitative Research Journal* (Vol. 9, Issue 2). doi:10.3316/qrij0902027
- Harsch, C. (2018). How Suitable Is the CEFR for Setting University Entrance Standards? *Language Assessment Quarterly*, 15(1), 102–108. <https://doi.org/10.1080/15434303.2017.1420793>
- Kamal, H. (2025). Teaching Arabic Today: Challenges, Strategies, and Opportunities in Islamic Higher Education. *International Journal of Learning, Teaching and Educational Research*, 24(10), 644–659. <https://doi.org/10.26803/ijlter.24.10.31>
- Koto, F., Li, H., Shatnawi, S., & Doughman, J. (2024). *ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic*. 4, 5622–5640.

- Lichtman, K., & VanPatten, B. (2021). Was Krashen right? Forty years later. *Foreign Language Annals*, 54(2), 283–305. <https://doi.org/10.1111/flan.12552>
- Liskin-Gasparro, J. E. (1984). The ACTFL Proficiency Guidelines: Gateway to Testing and Curriculum. *Foreign Language Annals*, 17(5), 475–489. <https://doi.org/10.1111/j.1944-9720.1984.tb01736.x>
- Llurda, E. (2000). On competence, proficiency, and communicative language ability. *International Journal of Applied Linguistics (United Kingdom)*, 10(1), 85–96. <https://doi.org/10.1111/j.1473-4192.2000.tb00141.x>
- Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample Size in Qualitative Interview Studies: Guided by Information Power. *Qualitative Health Research*, 26(13), 1753–1760. <https://doi.org/10.1177/1049732315617444>
- Maulani, H., Muthmainah, N., Khalid, S. M., Saleh, N., & Taufik, I. H. (2024). Investigation of the Reference Level Description for Arabic Proficiency Tests in Indonesia. *Jurnal Al Bayan: Jurnal Jurusan Pendidikan Bahasa Arab*, 16(1), 1. <https://doi.org/10.24042/albayan.v16i1.21566>
- McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14(1), 1–43. https://doi.org/10.1207/s1532690xc1401_1
- Miqdarul Khoir Syarofit, Hanik Mahliatussikah, Muhammad Alfian, & Salma, R. F. (2025). Analysing the Placement Test Questions at University in Egypt Based on the Common European Framework of Reference for Languages (CEFR). *Arabi : Journal of Arabic Studies*, 10(2), 121–136. <https://doi.org/10.24865/ajas.v10i2.984>
- Mohamed, S. (2023). The development of an Arabic curriculum framework based on a compilation of salient features from CEFR level descriptors. *Language Learning Journal*, 51(1), 33–47. <https://doi.org/10.1080/09571736.2021.1923781>
- Musthofa, T. (2022). CEFR-Based Policy in Arabic Language Teaching and Cultural Dimension in Indonesian Islamic Higher Education. *Eurasian Journal of Applied Linguistics*, 8(2), 96–107. <https://doi.org/10.32601/ejal.911545>
- North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System*, 23(4), 445–465. [https://doi.org/10.1016/0346-251X\(95\)00032-F](https://doi.org/10.1016/0346-251X(95)00032-F)
- Nurdianto, T., Hidayat, Y., & Fitrianto, I. A. A. (2025). Design of CEFR-Based Arabic Language Proficiency Test Indicators In The Arabic Language Education Department. *Ijaz Arabi Journal of Arabic Learning*, 8(3), 1182–1192. <https://doi.org/10.18860/ijazarabi.v8i3.28795>
- Oliver, B. (2013). Graduate attributes as a focus for institution-wide curriculum renewal: Innovations and challenges. *Higher Education Research and Development*, 32(3), 450–463. <https://doi.org/10.1080/07294360.2012.682052>
- R. Taufiqurrochman, & Lubna Farah. (2024). Transitioning from Monodisciplinary in Arabic Education: Indonesian Insights. *Evolutionary Studies in Imaginative Culture*, 915–924. <https://doi.org/10.70082/esiculture.vi.745>
- 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>
- Reay, D. (2018). The power of tests. In *Understanding Learning and Motivation in Youth*. <https://doi.org/10.4324/9781315170039-7>
- Roevers, C., & McNamara, T. (2006). Language testing: The social dimension. *International*

- Journal of Applied Linguistics*, 16(2), 242-258. <https://doi.org/10.1111/j.1473-4192.2006.00117.x>
- Thelwall, M., & Nevill, T. (2021). Is research with qualitative data more prevalent and impactful now? Interviews, case studies, focus groups and ethnographies. *Library and Information Science Research*, 43(2), 101094. <https://doi.org/10.1016/j.lisr.2021.101094>
- Tomaszewski, L. E., Zarestky, J., & Gonzalez, E. (2020). Planning Qualitative Research: Design and Decision Making for New Researchers. *International Journal of Qualitative Methods*, 19, 1-7. <https://doi.org/10.1177/1609406920967174>
- Verner, J. M., & Abdullah, L. M. (2012). Exploratory case study research: Outsourced project failure. *Information and Software Technology*, 54(8), 866-886. <https://doi.org/10.1016/j.infsof.2011.11.001>
- Wahyu Pratama, A., Mahliatussikhah, H., & Murtadho, N. (2024). Opportunities and Challenges of Common European Framework of Reference for Language (CEFR) Standardization in Arabic Language Learning. *Journal of Arabic Studies*, 9(1), 33-42. <http://dx.doi.org/10.24865/ajas.v9i1.774>
- Wang, H.-P., Kuo, B.-C., Chao, R.-C., & Tsai, Y.-H. (2012). The Development and Evaluation of a Computerized Adaptive Testing System for Chinese Proficiency - base on CEFR. *Procedia - Social and Behavioral Sciences*, 64, 34-42. <https://doi.org/10.1016/j.sbspro.2012.11.005>
- Wesolowski, B. C., & Wind, S. A. (2019). Pedagogical Considerations for Examining Rater Variability in Rater-Mediated Assessments: A Three-Model Framework. *Journal of Educational Measurement*, 56(3), 521-546. <https://doi.org/10.1111/jedm.12224>
- Wolff, D., & Costa, P. I. De. (2017). Expanding the language teacher identity landscape: An investigation of the emotions and strategies of a NNEST. *The Modern Language Journal*. <https://doi.org/10.1111/modl.12370>
- Wu, W. C. V, Hsieh, J. S. C., & Yang, J. C. (2017). Creating an online learning community in a flipped classroom to enhance EFL learners' oral proficiency. *Journal of Educational Technology &* <https://www.jstor.org/stable/90002170>
- Yonkers, N. Y. (1996). National standards in foreign language education project. *Standards for Foreign Language Learning in the 21st Century*, 37.