



Arabic Language Learning Assessment in Schools and Higher Education: A Scoping Review of Instruments, Skills, Digital Practices, and Evidence Gaps (2021–2026)

¹Abdul Basid, ²Nur Faizin, ³Habib Muharrom Sudarmawan, ⁴Moh. Fauzan, ⁵Abdul Adzim, ⁶Ahmad Munjin Nasih, ⁷Ali Ma'sum, ⁸Lilik Nur Kholidah, ⁹Achmad Sultoni, ¹⁰Mochammad Rizal Ramadhan, ¹¹Nur Mutamakkin, ¹²Mohammad Ahsanuddin, ¹³Nur Anisah Ridwan, ¹⁴Faris Khoirul Anam, ¹⁵Fery Fauzi, ¹⁶Muhammad Lukman Arifianto, ¹⁷Titis Thoriquityas

Universitas Negeri Malang, Indonesia

Corresponding Author: abdul.basid.fs@um.ac.id

Abstract

This article maps recent research on assessing Arabic as an additional language in schools and higher education. Using a Population–Concept–Context framework informed by PRISMA-ScR, a targeted bibliographic discovery process identified 40 candidate primary studies published in 2021–2026. These records address language-skill testing, writing rubrics, item analyses, formative feedback, proficiency descriptors, and digital assessment tools. Preliminary evidence suggests that expert approval and platform usability are frequently reported separately from score reliability, construct representation, and fairness. Psychometric studies using factor analysis, item response theory, and differential item functioning offer complementary evidence but do not establish universal validity across contexts. Because database-level searching, independent screening, and full-text adjudication have not been completed, the 40 records must not be described as a final included-study corpus. The article identifies methodological priorities for transparent, fair, and skill-sensitive Arabic assessment research.

Keywords: Arabic language assessment; language proficiency; digital assessment; scoping review

مستخلص البحث

تهدف هذه الدراسة إلى رسم خريطة أولية للأبحاث المتعلقة بتقويم تعلم اللغة العربية في المدارس ومؤسسات التعليم العالي. اعتمدت الدراسة إطار السكان والمفهوم والسياق، مع الاسترشاد بإرشادات المراجعة الاستطلاعية PRISMA-ScR، وحددت أربعين دراسة أولية مرشحة نُشرت بين عامي 2021 و2026. تشمل الموضوعات أدوات قياس المهارات اللغوية، وسلام تقدير الكتابة، وتحليل فقرات الاختبارات، والتقويم التكويني، ومستويات الكفاية، والمنصات الرقمية. وتوضح الخريطة الأولية ضرورة التمييز بين سهولة استخدام الوسائل الرقمية وصلاحيّة تفسير الدرجات وثباتها وعدالتها. كما تظهر الحاجة إلى توثيق إجراءات التصحيح والتوافق بين المقيمين والفروق بين مجموعات المتعلمين. ولم تستكمل بعد عمليات البحث الشامل في قواعد البيانات والفرز المستقل وفحص النصوص الكاملة، ولذلك لا تمثل الدراسات الأربعون حصيلة نهائية للمراجعة. وتخلص الدراسة إلى اقتراح أولويات بحثية لتحسين جودة تقويم اللغة العربية وشفافيته وشموليته. الكلمات الأساسية: تقويم اللغة العربية؛ الكفاية اللغوية؛ التقويم الرقمي؛ المراجعة الاستطلاعية

1. Introduction

1.1 Background and significance

Assessment is integral to Arabic language education because it shapes the inferences teachers, institutions, and learners draw about progress, placement, and achievement. Yet “evaluation” can denote substantially different activities: analysing the difficulty of a multiple-choice vocabulary item, scoring a learner’s oral interview, validating a writing rubric, evaluating a classroom application, or judging the consequences of a school examination. Treating these as a single evidence class obscures the distinct warrants required for each kind of conclusion (Messick, 1995).

Arabic presents assessment-specific design questions beyond those common to language education generally. Language scripts and optional vocalization, complex morphosyntax, possible differences between formal and colloquial registers, learner familiarity with reading and listening input, and the interaction of spelling with writing fluency all affect how tasks represent the intended construct. A vocabulary recognition test and an extended interaction task cannot be regarded as interchangeable proxies for global communicative competence merely because both are labelled “Arabic language tests.”

1.2 Related empirical research

Recent studies show why this distinction matters. Syihabuddin et al. (2021) reported empirical trialling of an internet-based assessment with 140 items, while Imaduddin et al. (2022) examined validity and internal consistency of a small school test. By

contrast, Boeriswati et al. (2023) operationalized academic writing criteria through a rubric, Yasin and Tarauni (2023) developed a CEFR-oriented speaking test, and Maulani et al. (2024) examined reference-level descriptors for Arabic proficiency tests. These projects approach language ability through different intended interpretations rather than through a single test design.

A parallel literature develops digital assessment media. Al Munawaroh (2021) and Mufti et al. (2022) examined Quizizz and related interfaces in classroom settings, while Rahmawati et al. (2023), Nurlaila et al. (2025), and Anawiyah and Qomariyah (2026) reported development or expert feasibility of further platforms. Such contributions are relevant to implementation, but high feasibility ratings do not by themselves show that proficiency scores are valid, equitable, or responsive to meaningful language development.

1.3 Research gap and rationale

Psychometric investigations have begun to examine those deeper questions. Bakti et al. (2025) reported differential item functioning across demographic and school groups; Subando et al. (2025) used multidimensional item response theory to diagnose item discrimination; and Pujiartha et al. (2025) reported exploratory and confirmatory evidence for an Arabic writing instrument. These studies suggest a developing but uneven assessment evidence base, a pattern more appropriately mapped than pooled into a single effectiveness estimate.

The rationale for a scoping review is therefore not to assert which application “works best,” but to map what is assessed, how constructs are operationalized, which validation claims are tested, where studies take place, and where the evidence is missing. Following Arksey and O’Malley (2005), Levac et al. (2010), Munn et al. (2018), Peters et al. (2024), and PRISMA-ScR guidance (Tricco et al., 2018), this article defines a transparent mapping question rather than a comparative meta-analysis.

1.4 Research questions and contributions

Research question: How has primary research conducted between 2021 and 10 October 2026 developed, implemented, and evaluated Arabic language learning assessments in formal educational contexts, and what evidence exists concerning language-skill coverage, score validity, reliability, fairness, technological integration, and consequences? The objectives are to (a) describe learner settings and language domains; (b) categorize instruments and study designs; (c) compare

the strength and kinds of evidence used to support assessment claims; and (d) identify actionable methodological gaps.

1.5 Organization of the article

The article first describes scope and review procedures, then organizes the candidate evidence into assessment-instrument and implementation charts. Subsequent sections discuss construct representation, scoring quality, digital applications, fairness, and consequential use before identifying the additional steps required for a reproducible PRISMA-ScR review.

2. Methods

2.1 Review design and protocol boundaries

This manuscript applies scoping-review logic to the development of an evidence chart and a protocol-ready narrative. A final scoping review requires documented database runs, duplicate removal, independent relevance screening, retrieval and full-text adjudication, and a reproducible account of exclusion reasons. Those procedures have not yet been completed. The 40 studies in the present tables are a targeted and provisionally charted seed corpus, not a numerically verified PRISMA-ScR eligible set. This distinction is retained explicitly in Figure 1.

2.2 Population–Concept–Context (PCC) and eligibility

Population: school or tertiary students learning Arabic in formal courses, teachers and assessment developers where their decisions concern learner assessment, and Arabic examination data. Concept: Arabic-language assessment instruments, performance measures, test evaluation, feedback, formative or summative scoring, proficiency-level descriptions, validity, reliability, fairness, washback, and digital assessment tools. Context: school, madrasa, Islamic boarding school, language institute, or higher-education Arabic courses internationally; studies may be reported in Arabic, English, or Indonesian.

For the target final corpus, eligible sources should be peer-reviewed original studies (empirical learner or teacher studies, psychometric analyses, original curriculum/test artefact analyses, and evaluated instrument-development projects) published 1 January 2021 to 10 October 2026. Exclude purely general Arabic pedagogy papers without an assessment question, systematic and narrative reviews, theoretical position papers without original chartable material, non-scholarly promotional content, assessments of unrelated subjects written in Arabic, and duplicate reports of the same sample unless findings are substantively distinct.

Papers with only instrument design and no field test must be flagged separately, not falsely treated as empirically validated instruments.

2.3 Search plan and discovery already performed

Proposed Scopus TITLE-ABS-KEY Boolean string: ("Arabic language" OR "Arabic as a foreign language" OR "Arabic as a second language" OR "maharah" OR "mahārah" OR "TOAFL") AND (assess* OR evaluat* OR test* OR psychometric* OR "item analysis" OR rubric* OR "formative feedback" OR "digital assessment") AND (learn* OR student* OR school* OR universit* OR "language proficiency") AND PUBYEAR > 2020 AND PUBYEAR < 2027. The comparable Web of Science Topic field should use TS=(...) with equivalent term groups and a 2021–2026 year filter. Arabic-script terms for language skills and assessment should be appended and manually tested for additional recall.

Planned source coverage: Scopus, Web of Science Core Collection, ERIC, DOAJ, Crossref-linked publisher pages, Arabic education journal portals, and handsearching of references in included studies. A supplementary title/abstract search should use (Arabic OR "تعليم العربية") AND ("assessment" OR "التقويم" OR "التقييم" OR "الاختبار") AND ("speaking" OR "reading" OR "writing" OR "listening" OR "مهاراة"). The final search log must record each database, date, exact query, returned count, export format, and duplicate-removal method.

In the current preparatory phase, web-indexed journal and publisher pages were interrogated using multiple English and Indonesian assessment-related strings. Forty publisher-linked primary-study candidates were retained in the preparatory chart after bibliographic matching, making 40 distinct charting records. This is purposeful discovery, not exhaustive database searching: it provides no defensible population denominator or database-specific record totals. The citation/DOI in Appendix A allows each charted study to be rechecked; lack of full-text evaluation is explicitly recorded.

2.4 Screening, charting, and synthesis plan

At least two reviewers should independently screen titles and abstracts against the PCC criteria after DOI-first deduplication, then assess retrieved full texts. Discrepancies should be resolved by discussion or a third reviewer. A reference manager should retain DOI, journal, year, language, country, study design, participant/sample size, assessment purpose, targeted skill and register, item or rubric form, type of scoring, psychometric parameters, digital environment, main

outcomes, and possible bias. Exclusion reasons must be logged and auditable. The preliminary chart here extracts only publisher-supported information; no unseen sample sizes or inferential statistics have been imputed.

Synthesis will group studies descriptively by assessment object rather than estimate pooled intervention effects. The planned mapping distinguishes (1) task/construct representation, (2) instrument development and standard setting, (3) score-interpretation evidence including reliability and fairness, and (4) consequences including formative feedback and test washback. Critical appraisal is optional under scoping methods but should be considered, with an explicit rubric, when drawing practical conclusions about high-stakes proficiency decisions.

2.5 PRISMA-ScR flow accounting

Figure 1 follows the identification, screening, eligibility and inclusion sequence of the supplied PRISMA template, while making the current incompleteness transparent. The 40 individually located study records in this manuscript cannot be equated with all records retrieved from Scopus or Web of Science, and no artificial exclusion, duplicate, or included counts are supplied. Upon completion of the database search, the figure should be replaced with the full PRISMA-ScR diagram reporting independent search yields and justified exclusions.

Figure 1. PRISMA-ScR workflow status (draft accounting)

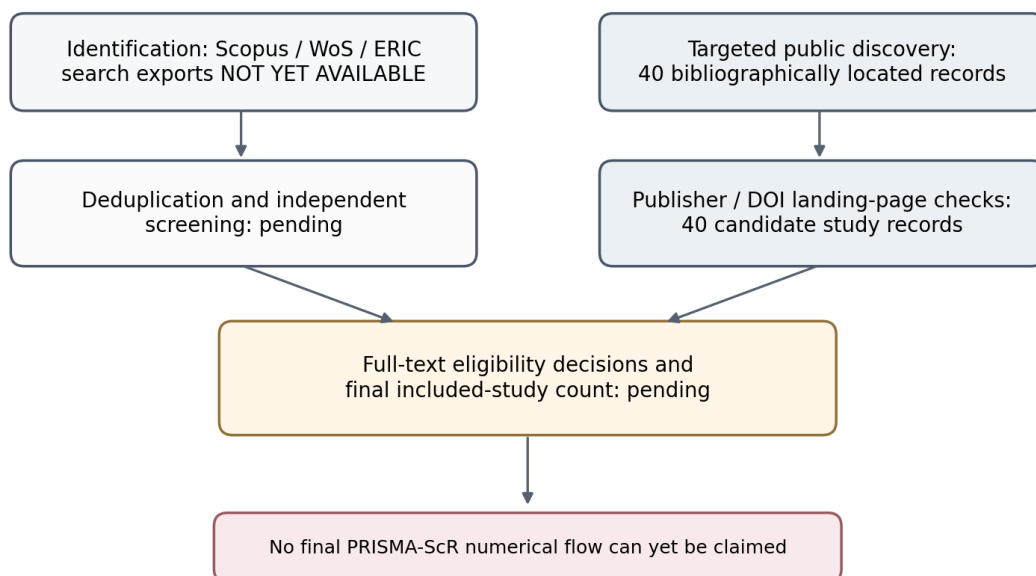


Figure 1. Draft PRISMA-ScR workflow status; unfilled stages must be replaced with audited counts after database export and full-text eligibility assessment.

3. Preliminary Evidence Map

3.1 Profile of the charted literature

The targeted chart contains 40 distinct candidate records dated 2021–2026, provisionally located and charted from journal-facing sources. By the descriptive categories assigned for this draft, 26 records focus primarily on test development, item analysis, skill-specific rubrics or assessment artefacts, and 14 focus principally on digital tools, classroom or institutional assessment practice. These categories represent an analytic coding decision, not a validated prevalence estimate for all Arabic-language assessment research. The candidate-year distribution is: 2021 ($n = 3$), 2022 ($n = 4$), 2023 ($n = 6$), 2024 ($n = 11$), 2025 ($n = 12$), 2026 ($n = 4$).

Table 1 charts the evidence closest to assessment constructs, tasks, rubrics, item characteristics and reference levels. Table 2 charts platform or classroom implementation, with attention to the gap between feasibility and measurement evidence. The tables deliberately give one row per primary study so that the discussion can be traced to named documents. Since studies were located through targeted discovery rather than a complete bibliographic search, neither table supports a claim about worldwide proportions.

The four-language-skill framework is present unevenly. Several sources focus on reading (Mufti et al., 2022; Imawan et al., 2023; Siregar et al., 2025), others on writing (Boeriswati et al., 2023; Pujiartha et al., 2025; Alauddin et al., 2026) or speaking (Yul et al., 2022; Yasin & Tarauni, 2023; Nuraini et al., 2024). Integrated four-skill assessment is more often named than comparably validated across all modalities. Interpretation should separate scores for receptive recognition from productive language performance.

Table 1. Primary studies on instruments, language skills, item quality and validity

One row per provisionally charted study. Findings are restricted to publisher abstracts and accessible bibliographic descriptions; claims require full-text confirmation.

Study	Setting / population	Assessment approach	Evidence charted / interpretation
Maftuhati (2021)	Islamic junior secondary	Interviews and test-document analysis	Analyzes validity and relevance of Arabic multiple-choice final exam items
Syihabuddin (2021)	Online proficiency test	140 items developed; empirical pilot and revision	81 valid items, mean reliability 0.67 reported
Imaduddin (2022)	School; grade II	Ten MCQ; classical item analysis	Six of 10 items classified as valid; Cronbach alpha 0.82
Ridlo (2022)	Islamic boarding school	Case study of test and non-test evaluation	Documents multiple evaluation forms; outcome claims are not causal evidence
Yul (2022)	National school final examination	Critical document/policy analysis	Speaking was not represented in the assessed final-exam format
Boeriswati (2023)	University writing	Rubric development; mixed-method design	Operationalizes academic writing against qualification descriptors
Imawan (2023)	Textbook; receptive skills	Primary textbook document analysis	Maps missing receptive assessment variety and proposes forms

Study	Setting / population	Assessment approach	Evidence charted / interpretation
Munawarah (2023)	Secondary textbook; reading	Textbook assessment analysis	Identifies need for HOTS-oriented reading assessment tasks
Yasin (2023)	University speaking	B1 CEFR-oriented speaking-test development	Reports item validity, reliability, difficulty and discrimination analyses
Ahmad (2024)	Junior secondary; Indonesian school	Document analysis of grammar and vocabulary test items	Finds context-bound and discrete-item balance in midterm examination
Amelia (2024)	Elementary school	Qualitative interviews and document analysis	Describes use of criterion-referenced decisions in classroom assessment
Kadir (2024)	School; 101 respondents	Mixed qualitative and quantitative item analysis	Maps discrimination, difficulty and higher-order-thinking demand
Makfiro (2024)	Islamic boarding school; grammar	Qualitative case study, portfolio artifacts	Portfolios track learner grammar work and process
Maulani (2024)	CEFR-adapted proficiency	Reference-level description development	Maps alignment and adaptation of Arabic proficiency descriptors
Nuraini (2024)	Elementary speaking	Qualitative formative assessment investigation	Uses formative observations to trace oral-learning progress
Pangestika (2024)	Grade VII textbook	Primary textbook content analysis	Identifies and redesigns discrete-point assessment approach

Study	Setting / population	Assessment approach	Evidence charted / interpretation
Tanjung (2024)	Islamic junior secondary	Item validity, reliability, discrimination, difficulty study	Evaluates characteristics of school end-of-year examination items
Alhazmi (2025)	University; 20 international students	Application of processability framework	Maps learner-level proficiency through production criteria
Arifa (2025)	University placement/proficiency	Web test prototype; R&D	150 MCQs across listening, structure/writing knowledge, reading
Azmy (2025)	Grade VII textbook	Affective instrument development	Proposes affective assessment linked to communicative textbook tasks
Bakti (2025)	25 schools; 1,157 students	IRT and differential item functioning	DIF detected across gender, study specialization and geography
Muslimah (2025)	Islamic boarding school; writing	Case study of summative writing assessment	Documents enacted writing assessment practices and setting constraints
Pujiartha (2025)	Higher education; 248 learners	Exploratory and confirmatory factor analyses	Four-domain Arabic writing instrument with reported fit and reliability
Siregar (2025)	Reading-material analysis	Document-based instrument development	Proposes pronunciation, selected-response and essay tasks; no field psychometric trial
Subando (2025)	Grade XII; 321 students	MIRT 2PL and exploratory factor analysis	Sixteen items had negative discrimination; ease dominated test items

Study	Setting / population	Assessment approach	Evidence charted / interpretation
Alauddin (2026)	Senior secondary; writing	Qualitative writing assessment study	Identifies emphasis on accuracy and the need for analytic rubric

3.2 Forms of measurement evidence

Classical item analyses expose limitations that would be hidden by a simple percentage score. Imaduddin et al. (2022) reported that six of ten questions satisfied their validity criterion, with Cronbach's alpha of 0.82 for the sampled test. In an internet-based proficiency development study, Syihabuddin et al. (2021) found 81 of 140 questions valid and reported reliability of 0.67. These values refer to different tasks and populations: they cannot be compared as though the larger alpha necessarily indicates better construct representation, nor do either sets of statistics establish fairness for all learners.

Later measurement models address richer questions. Subando et al. (2025) reported 16 negatively discriminating items in a sample of 321 students and documented an excess of easy items, calling into question the targeting of the test to stronger candidates. Bakti et al. (2025), using response data for 1,157 students from 25 schools, reported subgroup differential item functioning, including 11 items by gender, 23 by specialization and 37 by geographic region. Those counts overlap in their possible item memberships; they should not be added into a number of unique biased items. Table 1 thus differentiates internal-consistency, structural, item-functioning and fairness evidence rather than collapsing them into a single “validity” category.

Writing assessment offers a contrasting object of inference. Boeriswati et al. (2023) developed a rubric tied to Indonesian qualifications criteria; Pujiartha et al. (2025) used data from 248 students to examine four dimensions—ideas/content, structure/organization, style/rhetoric and language/grammar—and reported satisfactory factor structure. The design and scoring of extended written texts require rater guidance, moderation, and agreement estimates in addition to internal consistency. Such details should be systematically retrieved before making claims about reliability of scoring across institutions.

Reference-level and construct alignment require equal attention. Yasin and Tarauni (2023) designed a CEFR B1-oriented speaking test, while Maulani et al. (2024) investigated adapted proficiency reference descriptions. Arifa et al. (2025) described a 150-item web-based proficiency prototype that emphasizes listening, structural knowledge and reading, whereas Yul et al. (2022) criticized the absence of speaking from a high-stakes school test. The juxtaposition points to a recurrent gap: declarative proficiency labels should be matched to actual task coverage, rating conditions and interpretation of performance.

Textbook analyses and practice cases, including Munawarah and Baroroh (2023), Pangestika et al. (2024), Kadir et al. (2024), Tanjung et al. (2024), and Siregar et al. (2025), help identify misalignments between intended higher-order or receptive constructs and the assessment questions offered. However, design recommendations based on content analysis must be kept distinct from field validation. A test blueprint may improve the plausibility of content representation while leaving score precision, criterion interpretation and stakeholder consequences untested.

3.3 Digital tools and assessment practice

Digital assessment studies make a different contribution. Al Munawaroh (2021) describes the operation of Quizizz in a school setting, and Mufti et al. (2022) reports learners' perceived advantages and challenges across Socrative, Quizizz and Google Form, including internet access and limitations in Arabic font display. Rahmawati et al. (2023) develops a Quizizz-based evaluation tool, and Nur et al. (2023) uses an ADDIE process for a Kahoot-based formative vocabulary assessment. These sources primarily concern feasibility, classroom perceptions and development processes; none should be read as a general proof of gains in Arabic proficiency merely from choosing a named platform.

Not all development studies use the same design: Nur et al. (2023) also reports an experimental-versus-control comparison for Kahoot-based vocabulary assessment. Even where group contrasts are reported, their generalizability and the interpretation of language-score differences require attention to the measures and instructional settings used.

The targeted record set includes both empirical and prototype-focused applications. Nugraha et al. (2025) reported a normalized gain of 59% and a mean student-response percentage of 78.9% in a one-group pre/post design, which lacks the counterfactual required for a platform-specific causal attribution. Nurlaila et al. (2025) reported expert/user feasibility and practicality means of 4.26 and 4.395; Anawiyah and Qomariyah (2026) reported material, media, and language expert validation percentages of 75%, 87%, and 81%. Those metrics characterize validation within their development frameworks; they do not establish an independent psychometric validity argument for language scores.

Table 2 records the narrower evidence each digital or classroom source directly supports. For a more mature literature, subsequent studies should compare electronic and paper formats for measurement invariance, document input methods

for Arabic script and diacritics, assess accommodations for learners with differing technological access, and report the role of teachers in confirming automated feedback (Rizki et al., 2024).

Table 2. Primary studies on digital interfaces, formative assessment and implementation

One row per provisionally charted primary record; 'positive response' indicates participant or expert feedback, not causal learning effectiveness.

Study	Educational context	Tool / research method	Directly supported evidence
Al Munawaroh (2021)	High school	Descriptive qualitative teacher and student study	Quizizz eased teacher processing and received favorable student responses
Mufti (2022)	University reading course	Student perception questionnaire	Ease and motivation reported; internet and Arabic fonts caused challenges
Nur (2023)	Vocational school; vocabulary	ADDIE research-and-development	Designs Kahoot formative instrument for Arabic vocabulary
Rahmawati (2023)	Islamic junior secondary	Evaluation tool development and testing	Reports validity, practicality and effectiveness aims for Quizizz format
Ahmad (2024)	Boarding school; oral and written tests	Institutional test case study	Describes oral/written testing alignment with school curriculum
Chalista (2024)	Secondary school; reading/listening	Digital assessment tool development	Reports feasibility and practicality of a quiz assessment application
Rizki (2024)	150 participants; reading	Descriptive design with classroom field component	Proposes human review of AI feedback for six-stage reading assessment
Jamhuri (2025)	Islamic senior secondary	ADDIE design and expert/user trials	Reports content and media validation plus classroom practicality

Study	Educational context	Tool / research method	Directly supported evidence
Nugraha (2025)	Elementary school	One-group pre/post design	Reports normalized gain 59% and feedback mean 78.9%
Nurlaila (2025)	Islamic senior secondary	R&D expert and user validation	Feasibility mean 4.26; practicality mean 4.395
Robbani (2025)	High school; 92 participants	Application-based exam washback study	Investigates consequences of digital testing for learner behavior
Anawiyah (2026)	Islamic senior secondary	ADD-phase R&D expert validation	Material 75%, media 87%, language 81% expert ratings; learning gains untested
Ma'arif (2026)	Multiple skills	Case study of digital assessment practice	Differentiates digital assessment affordances across four skills
Sani (2026)	Grade VII curriculum	Assessment prototype based on textbook audit	Maps curriculum alignment and diagnostic gaps in textbook tests

3.4 A construct-to-consequence map

Figure 2 formalizes a conceptual distinction emergent from Tables 1 and 2. First, assessments target different language skills or components; second, tasks and rubrics make these constructs observable; third, measurements require warrants such as structural coherence and subgroup comparability; and fourth, scores are used to supply feedback, placement, grading or policy decisions. This diagram is a proposal for synthesizing heterogeneous evidence, not evidence of a statistically validated causal sequence.

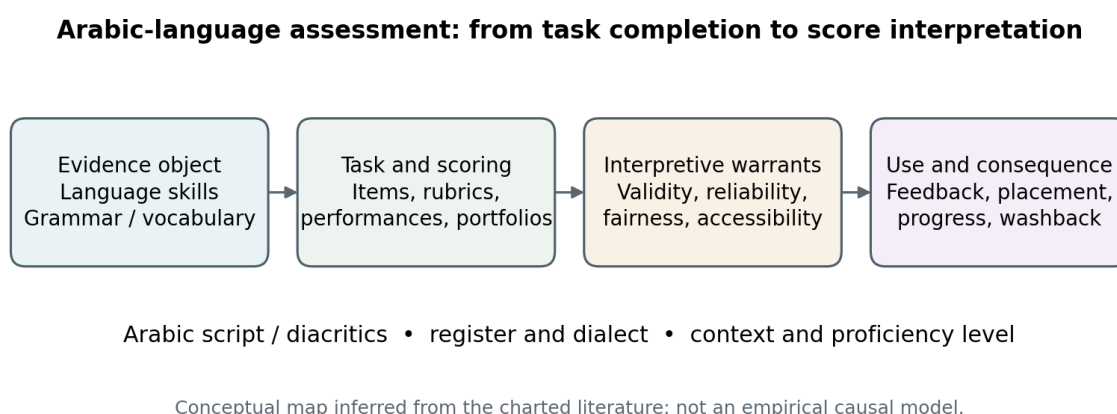


Figure 2. Conceptual linkage among Arabic-language constructs, tasks, score interpretation and assessment consequences; derived as an analytic framework for this review.

The analytic classifications also separate evidence on measurement properties from evidence on classroom implementation. Their descriptive counts are features of the seed corpus, and any stronger statement about prevalence or geographic distribution awaits complete database retrieval and eligibility screening.

Figure 1 identifies which review stages remain unaudited; the contrast between the instrument studies in Table 1 and practice studies in Table 2 is synthesized below rather than treated as a pooled estimate.

4. Discussion

4.1 The central argument: technology does not substitute for construct validity

The provisional map indicates a measurement–implementation gap. Technology-focused reports frequently address ease of deployment, perceived engagement, or experts' judgments of educational media. More demanding psychometric studies address the interpretation of scores, item functioning, and dimensionality. These

bodies of work are complementary, but the underlying evidence standards differ: a program can be enjoyable and administratively efficient while still giving unstable or systematically biased information about Arabic proficiency.

Consider a high-stakes placement decision. A quiz that functions smoothly in class might show minimal evidence that its items represent a communicative domain, discriminate across the intended ability range, remain invariant across learners, or connect to an externally meaningful proficiency description. The developmental studies in Table 2 do not necessarily make such claims; the editorial error would be to assume those outcomes from their feasibility scores. By comparison, Bakti et al. (2025), Subando et al. (2025), and Pujiartha et al. (2025) illustrate types of psychometric evidence that matter for defensible score use, although their specific findings require context-specific interpretation.

This distinction motivates a shift in the language used to discuss “effectiveness.” Unless a study has a suitable comparison and measurement design, the defensible conclusion is often that a particular tool was developed, rated usable, or associated with changes under observed conditions. This careful wording improves the integrity of the synthesis while protecting the diverse local case studies from unfair methodological expectations that were never part of their research aims.

4.2 Language domains, register and representation

Arabic assessment must differentiate linguistic knowledge from meaningful communicative performance. Morphological and syntactic discrimination may have a legitimate place in diagnosis, but inferring oral fluency from grammar MCQs is a construct-coverage error. The concern articulated by Yul et al. (2022) is especially salient for examinations that claim four-skill coverage without directly sampling oral production. Speaking requires scoring criteria for interaction, intelligibility, discourse, task completion and, where relevant, sociolinguistic appropriateness.

Writing likewise calls for analytic treatment of content, organization, rhetorical effect, and language form. The rubric-development and factor-analysis efforts of Boeriswati et al. (2023) and Pujiartha et al. (2025) provide a useful starting point. Research should identify whether raters were trained, whether scores reflect differences among raters, whether tasks invite sufficient variability in performance, and how orthographic or script conventions are handled for non-native learners. When constructing a scoring system, researchers should state clearly whether the test represents Modern Standard Arabic, formal academic writing, classroom Arabic, or conversational varieties.

Receptive assessment demands more than countable correct answers. Reading tasks based on textbook material can overemphasize literal recall, and listening tasks can confound auditory discrimination with vocabulary or background knowledge. Primary artefact analyses in Table 1 identify variation in question types; future validation should test passage difficulty, audio conditions, distractor design, item-dependence, and learner cognitive processes. Comparing outcomes across language skills is premature where tasks differ radically in their breadth, scoring reliability and proficiency assumptions.

4.3 Equity, consequences and the interpretation of scores

Educational assessment should be evaluated for fairness as well as aggregate reliability. Bakti et al. (2025) provides an instructive example through its subgroup differential item-functioning analysis. Detection of DIF does not by itself prove why an item behaves differently, but it establishes a reason to examine content, contextual exposure, teaching opportunity, register, or language familiarity before allowing the affected scores to determine opportunity. Similar analyses could explore school type, prior Qur'anic Arabic exposure, first language, disability accommodations, and device-mediated literacy without treating subgroup membership as a deficit.

Consequences of assessment also merit direct inquiry. Robbani et al. (2025) considers washback from an application-based examination, while formative assessment work (Nuraini et al., 2024; Makfiro et al., 2024) emphasizes how teacher observations or portfolios trace learning processes. Reporting improved participation or greater engagement is relevant, but should not substitute for evidence about feedback use, changed learner strategy, improvement on independent measures, or educational decisions. In complex multilingual classrooms, consequences may be as important as initial test design.

Ethical development of automated systems is another priority. A human-supervised design such as Rizki et al. (2024) anticipates the importance of teacher interpretation, but systematic trials are required to determine when algorithm-generated judgments are accurate and for whom. Arabic diacritization, handwriting, standard versus colloquial input, spelling correction, speech recognition, and data privacy could each create distinctive validity challenges; the charted studies do not yet license generalized claims that an AI system has resolved these problems.

4.4 The concentration of evidence and its limits

The study locations readily found by targeted discovery are strongly concentrated in Indonesian school, madrasa, pesantren and university settings. This apparent concentration is a property of the located evidence set, not a credible estimate of global publication patterns. A reproducible bibliographic search in Arabic script and additional regional scholarly infrastructures may reveal substantial work in the Middle East, North Africa and other Asian settings. Until such retrieval is undertaken, geographic absence must be described as a gap in search coverage rather than an absence of research.

Sources also vary in what “validation” means. For some R&D studies, validation refers to expert appraisal of content and media usability; for item analysis, it refers to classical indices; for psychometric models, it refers to structural or subgroup response patterns. Consequently, the term should never be interpreted without its associated data and analysis. This is the most important methodological caution of the current review: the apparent growth of evaluation technology cannot be equated with a demonstrated rise in measurement quality.

Figure 2 and Tables 1–2 point toward a practical research strategy: define the intended claim first, then match task design, scorer or algorithm evidence, population and score-use consequences to that claim. Without those links, favorable performance reports risk becoming demonstrations of platform acceptability rather than evidence about language assessment.

5. Research Priorities and Implications

Figure 3 proposes an evidence-maturity sequence that researchers and journal reviewers can use when interpreting future assessment studies. Level I concerns feasibility and expert review, Level II concerns the quality and stability of scores, Level III tests fairness and accessibility, and Level IV examines decisions and learning consequences. These are complementary evidence domains rather than a rigid sequence in which later stages mechanically invalidate earlier-stage design research.

Priority 1: Multidimensional proficiency measurement. Develop assessment blueprints explicitly spanning listening, speaking, reading and writing, and explain the intended register, learner level, item format and reporting scale. Link score meaning to observable performances and test comparability across administrations.

Priority 2: Stronger validation reports. Beyond expert ratings, report content mapping, respondent cognitive processes, rater training, inter-rater agreement,

item difficulty/discrimination, structural analyses where suitable, confidence intervals, and decisions implied by proposed cut scores.

Priority 3: Arabic-specific digital assessment. Test input and presentation of right-to-left script, variable diacritic support, accessibility for learners with limited devices, fidelity of audio delivery, automated speech or writing feedback against trained human judgments, and data governance.

Priority 4: Broader contexts and consequential evidence. Expand the country and educational-context sampling of the review itself; develop cross-site replication; assess fairness across linguistic and instructional backgrounds; and study whether feedback changes subsequent language use rather than only satisfaction ratings.

Priority 5: Reporting infrastructure. Archive anonymized blueprints, sample characteristics, scoring rubrics, item-analysis scripts, codebooks, and search logs where permitted. Such transparency makes individual studies more usable for a later targeted systematic review or an assessment-focused evidence-and-gap map.

Progression of evidence needed for defensible assessment claims

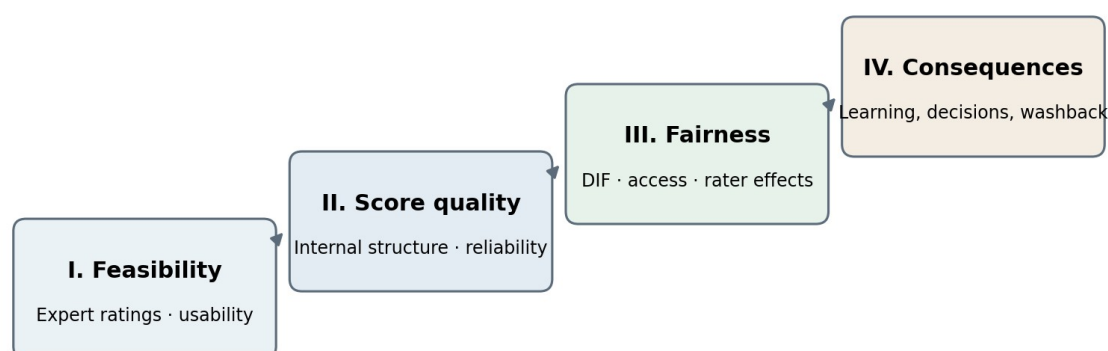


Figure 3. Proposed domains of evidence maturity for Arabic language assessment studies, from feasibility through score interpretation, fairness and consequences.

The progression illustrated in Figure 3 is therefore a research-planning heuristic, not a rating or ranking of the 40 candidate studies. Together with Figure 2, it shows why a credible argument about digital Arabic testing should connect tasks, score interpretation, fairness, and consequences.

6. Limitations and Next Review Steps

This manuscript is a substantive first draft, not a completed systematic search. Scopus, Web of Science, and ERIC exports were unavailable for auditable database counting; studies were located through targeted indexed discovery, publisher

pages, and DOI-level bibliographic checking. The 40 charted records represent a reproducible starting set of named candidate studies, not all records worldwide, and full-text review is incomplete. Some source descriptions may emphasize authors' own claims without independent quality appraisal. Selection therefore may privilege English-language indexing, Indonesian journals, and recent digitally visible records.

Before submission, reviewers must execute the dated database queries, preserve complete exports, remove duplicates, screen independently, retrieve all potentially eligible articles, document every exclusion at full-text stage, and replace Figure 1 with authentic PRISMA-ScR counts. Article titles, exact authorship, page spans, published dates and statistical claims should be checked against full texts and publisher/Crossref records. The working synthesis and any gaps should then be revised using the finalized eligible sample. Appendix A supplies the starting claim-source ledger and verification limitations; Appendix B identifies the principal publication risks.

7. Conclusion

The provisional research map suggests a diverse and increasingly technology-mediated landscape for assessing Arabic language learning across school and university contexts. The strongest conceptual lesson is that assessment media, assessment tasks, and score interpretations must not be conflated: the feasibility of a Quizizz-based classroom tool differs from evidence for the fairness or structural validity of a language examination. Recent studies of Arabic writing rubrics, multidimensional item analysis, and differential item functioning show promising directions for more defensible measurement, while formative, portfolio, and digital case studies reveal how assessment operates in authentic classrooms.

The emerging research agenda is to integrate language-domain representation, item and rater quality, fairness, accessibility, and the consequences of using scores. A fully reproducible PRISMA-ScR search and independent full-text screening are still necessary to establish the completeness of this map, calibrate claims about evidence gaps, and transform the draft into a journal-submission-ready scoping review.

References

- Ahmad, H., & Auliya, A. N. (2024). Analisis soal tes kompetensi kebahasaan pada penilaian tengah semester bahasa Arab sekolah menengah pertama. *Al Mi'yar: Jurnal Ilmiah Pembelajaran Bahasa Arab dan Kebahasaaraban*, 7(1). <https://doi.org/10.35931/am.v7i1.3157>
- Ahmad, F. S., Yasin, A., Sofian, A. A., Khalifah, N. F., & Rahmawati, N. (2024). Curriculum enhancement and language assessment in pesantren: Examining Arabic oral and written tests at Pondok Modern Darussalam Gontor Putri Campus 1. *AFKARINA: Jurnal Pendidikan Agama Islam*, 9(1), 71–82. <https://doi.org/10.33650/afkarina.v9i1.15528>
- Al Munawaroh, N. T. (2021). The use of Quizizz online software in the evaluation of Arabic learning. *ALSUNIYAT: Jurnal Penelitian Bahasa, Sastra, dan Budaya Arab*, 4(1), 29–46. <https://doi.org/10.17509/alsuniyat.v4i1.31258>
- Alauddin, I., Nurdin, N. I., Aqilah, N., & Baihaqi, M. (2026). A qualitative assessment of students' Arabic writing skills (Kitabah) at SMA Bina Taruna Surabaya. *Tsaqofiya: Jurnal Pendidikan Bahasa dan Sastra Arab*, 8(1). <https://doi.org/10.21154/tsaqofiya.v8i1.1067>
- Alhazmi, M. Z. (2025). Applied study of the processability theory of Arabic learners in the Islamic University of Madinah. *Ijaz Arabi: Journal of Arabic Learning*, 8(1). <https://doi.org/10.18860/ijazarabi.v8i1.29356>
- Amelia, F., Bahrin, M. F., & Qomari, N. (2024). Implementation of criterion-referenced assessment in Arabic language education Muslim Cendekia Malang Elementary School. *Abjadia: International Journal of Education*, 9(2). <https://doi.org/10.18860/abj.v9i2.28047>
- Anawiyah, A., & Qomariyah, L. (2026). Design and validation of a JotForm-based Arabic language assessment model to enhance higher-order thinking skills in Islamic senior high schools. *Arabiyatuna: Jurnal Bahasa Arab*, 10(1), 361–386. <https://doi.org/10.29240/jba.v10i1.16070>
- Arifa, F., Kamal, M., Tajuddin, S., Bahtiar, I. R., Masyitoh, S., & Yani, A. (2025). Development of a web-based Arabic competency test for Indonesian speakers as a reference for measuring Arabic proficiency levels of undergraduate and graduate students in Arabic language education at Universitas Negeri Jakarta. *Arabiyatuna: Jurnal Bahasa Arab*, 9(2), 651–666. <https://doi.org/10.29240/jba.v9i2.13241>

- Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19–32. <https://doi.org/10.1080/1364557032000119616>
- Azmy, M. U., Baroroh, R. U., Novian, Z., & Salsabila, S. (2025). Innovations in affective assessment in Arabic language books for grade VII based on communicative approach. *Arabiyatuna: Jurnal Bahasa Arab*, 9(2), 709–724. <https://doi.org/10.29240/jba.v9i2.13740>
- Bakti, A. A., Marzuki, M., Ibrahim, Z. S., Tuanaya, R., & binti Yacob, N. Y. (2025). Differential item functioning analysis of Arabic language exams across gender, study specialization, and geographic region in senior high schools. *REID (Research and Evaluation in Education)*, 11(1), 59–74. <https://doi.org/10.21831/reid.v11i1.85961>
- Boeriswati, E., Mayuni, I., Fahmi, A. K., & Surong, K. (2023). The development of assessment rubric of Arabic language writing based on National Qualifications Framework of Indonesia. *Jurnal Al Bayan: Jurnal Jurusan Pendidikan Bahasa Arab*, 15(1), 261–277. <https://doi.org/10.24042/albayan.v15i1.16302>
- Chalista, A., Kesuma, G. C., Amrulloh, M. A., Zuliana, E., & Wahidah, Y. L. (2024). Developing Arabic language learning assessment tools using the Competition application in secondary school. *ALSUNIYAT: Jurnal Penelitian Bahasa, Sastra, dan Budaya Arab*, 7(2). <https://doi.org/10.17509/alsuniyat.v7i2.69424>
- Imaduddin, M. F., Maulani, H., & Taufik, I. H. (2022). Test the validity and reliability of Arabic learning questions. *Arabi: Journal of Arabic Studies*, 7(2), 198–207. <https://doi.org/10.24865/ajas.v7i2.523>
- Imawan, Y. (2023). Inovasi bentuk penilaian tes kompetensi reseptif berbahasa Arab dalam buku *Al-'Arabiyyah Lin Nasyi'in Jilid 1*. *Tsaqofiya: Jurnal Pendidikan Bahasa dan Sastra Arab*, 5(1), 19–36. <https://doi.org/10.21154/tsaqofiya.v5i1.177>
- Jamhuri, W., & Kasih, B. W. N. (2025). Development Quizizz application as instrument evaluation in Arabic language learning at Madrasah Aliyah. *IJ-ATL: International Journal of Arabic Teaching and Learning*, 8(2). <https://doi.org/10.33650/ijat.v8i2.9696>
- Kadir, S. D., Sarif, S., & Fuadi, A. H. N. (2024). Item analysis of Arabic thematic questions to determine thinking level ability. *ELOQUENCE: Journal of Foreign Language*, 3(1), 28–41. <https://doi.org/10.58194/eloquence.v3i1.1498>

- Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, 5, 69. <https://doi.org/10.1186/1748-5908-5-69>
- Ma'arif, A. S., & Bakhiet, B. M. (2026). Implementing digital assessment in Arabic language learning: A case study of listening, speaking, reading, and writing skills. *Funun Arabiyyah: Journal of Arabic Education and Literature*, 2(2), 147–175. <https://doi.org/10.64131/fununarabiyyah.v2i2.146>
- Maftuhati, Musyafa'ah, N., & Maulidya, S. (2021). Tahlilu jaudati bunudi al-ikhtiyarat al-muta'addidah li-ta'lim 'anashiri al-lughah al-'arabiyyah. *Alibbaa': Jurnal Pendidikan Bahasa Arab*, 2(1). <https://doi.org/10.19105/ajpba.v2i1.4119>
- Makfiro, N., Hermiwati, H., Sumiati, S., & Muzadi, N. H. (2024). Evaluation of portfolio assessment of student work in Arabic language learning. *Tanwir Arabiyyah: Arabic as Foreign Language Journal*, 4(2). <https://doi.org/10.31869/afl.v4i2.6008>
- 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), 1–15. <https://doi.org/10.24042/albayan.v16i1.21566>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Mufti, A., Maksudin, M., Baroroh, R. U., & Setiyawan, A. (2022). Socrative, Quizizz, and Google Form as online-based evaluation tools for Maharah Al-Qira'ah learning. *Maharaat: Jurnal Pendidikan Bahasa Arab*, 4(2), 120–136. <https://doi.org/10.18196/mht.v4i2.14229>
- Munawarah, K., & Baroroh, R. U. (2023). Innovation of assessment techniques on Maharah Qiraah in Arabic textbooks based on higher order thinking skills (HOTS). *Jurnal Basicedu*, 7(4), 2143–2150. <https://doi.org/10.31004/basicedu.v7i4.5515>
- Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, 18, 143. <https://doi.org/10.1186/s12874-018-0611-x>

- Muslimah, M., Sulalah, I. A., Rohmawati, R., Sofi, N., & Zunairoh, Y. (2025). Best practice of summative assessment for Arabic writing proficiency: Evidence from Pondok Pesantren Modern Ar-Rahmah Kediri, Indonesia. *Asalibuna*, 8(2). <https://doi.org/10.30762/asalibuna.v8i02.5221>
- Nugraha, D., Komara, W. R., Odang, O., Ruhendi, A., & Maryani, N. (2025). Improving the quality of Arabic language learning assessment using the paper-based Quizizz mode. *Maharaat: Jurnal Pendidikan Bahasa Arab*, 8(1). <https://doi.org/10.18196/mht.v8i1.33384>
- Nur, A. M., Mudrofin, M., & Bukhori, N. F. (2023). Tatwir adawat al-taqwim al-takwini al-qa'imah ala Kahoot li-tarqiyat itqan al-mufradat fi ta'lim al-lughah al-arabiyyah bil-madrasah al-mihaniyyah. *Alsina: Journal of Arabic Studies*, 5(2). <https://doi.org/10.21580/alsina.5.2.22933>
- Nuraini, A., Jundi, M., Hasibuan, R., & Djola, F. (2024). Student progress tracing of Arabic learning: A qualitative investigation of speaking skills formative evaluation in Islamic elementary school. *Al-Ihda': Media Ilmiah Bahasa Arab*, 12(1), 1–12. <https://doi.org/10.58645/alihda.v12i1.520>
- Nurlaila, N., Aisyah, D., Al Husna, L., Anisa, A., Kamaluddin, K., Widayanti, R., & Asfar, H. (2025). Developing a web-based interactive Arabic assessment to stimulate students' interest in learning. *Ijaz Arabi: Journal of Arabic Learning*, 8(3). <https://doi.org/10.18860/ijazarabi.v8i3.29016>
- Pangestika, E., Baroroh, R. U., & Maknuna, S. L. (2024). Innovation discrete approach based on Arabic language test assessment. *Arabiyatuna: Jurnal Bahasa Arab*, 8(1), 395–410. <https://doi.org/10.29240/jba.v8i1.8992>
- Peters, M. D. J., Godfrey, C., McInerney, P., Munn, Z., Tricco, A. C., & Khalil, H. (2024). Scoping reviews. *JBIMES-24-09*. <https://doi.org/10.46658/JBIMES-24-09>
- Pujiartha, L. A., Ritonga, A. W., Desrani, A., & Supriatna, A. (2025). Developing an instrument to assess writing skills in Arabic as a second language: A confirmatory factor analysis approach. *Al-Ta'lim Journal*, 32(2). <https://doi.org/10.15548/jt.v32i2.937>
- Rahmawati, A. E., Zulhannan, Hijriyyah, U., Erlina, & Koderi (2023). Development of the Quizizz application-based evaluation tool for learning Arabic for MTs. *Mantiqu Tayr: Journal of Arabic Language*, 3(2), 135–150. <https://doi.org/10.25217/mantiquatayr.v3i2.3484>

- Ridlo, U. (2022). Multiple intelligences in evaluation of Arabic learning at Islamic boarding school. *Arabiyatuna: Jurnal Bahasa Arab*, 6(1). <https://doi.org/10.29240/jba.v6i1.3798>
- Rizki, R. B., Rizal, M. F., & Rahmawati, C. (2024). Design of integrating Qalam AI and language teacher competencies in evaluating Arabic reading skill. *Tanwir Arabiyyah: Arabic as Foreign Language Journal*, 4(2), 257–266. <https://doi.org/10.31869/aflj.v4i2.5930>
- Robbani, A. S., Mirma, A. F., Munadi, Y., Safa, B. S. S., & Badar, S. (2025). The washback effects of Arabic language final examination using the GeSchool application. *Jurnal Al Bayan: Jurnal Jurusan Pendidikan Bahasa Arab*, 17(1). <https://doi.org/10.24042/albayan.v17i1.25932>
- Sani, N., Rahayu, D., JH, V. A., Irhamudin, I., & Rahman, Z. (2026). Innovating assessment for grade VII Arabic curriculum: A design–development study of Ministry textbooks. *Lingua: Jurnal Keilmuan dan Kependidikan Bahasa Arab*, 12(1). <https://doi.org/10.32678/lingua.v12i1.12647>
- Siregar, A. F., Muhajir, Bayzura, S., Salsabila, K., Himam, A. S., & Alfiroh, N. (2025). Assessment instrument development for Maharah Qira'ah: A case study on Qira'ah Rasyidah book volume 5. *Jurnal Al Bayan: Jurnal Jurusan Pendidikan Bahasa Arab*, 17(1), 238–254. <https://doi.org/10.24042/3ntsqb89>
- Subando, J., Astoko, D. B., Amin, L. H., Budiharjo, & Embong, R. (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>
- Syihabuddin, Abdurrahman, M., Akmalia, F., & Abdussalam, A. (2021). Wikalah iBT: How is an internet-based Arabic language proficiency test instrument developed?. *Ijaz Arabi: Journal of Arabic Learning*, 4(3), 535–547. <https://doi.org/10.18860/ijazarabi.v4i3.12198>
- Tanjung, M. A. H. R., Fahmi, A. A., Rahmanita, F., Habibah, I. F., & Qomari, N. (2024). Analisis butir soal penilaian akhir tahun pelajaran Bahasa Arab kelas VII MTs Al-Ma'arif Rakit Banjarnegara Jawa Tengah. *Mantiqu Tayr: Journal of Arabic Language*, 4(1), 347–367. <https://doi.org/10.25217/mantikutayr.v4i1.4038>
- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., ... Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR):

Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467–473.
<https://doi.org/10.7326/M18-0850>

Yasin, A., & Tarauni, H. (2023). Tashmim al-Ikhtibar li Maharah al-Kalam al-Arabi Wafqhan li al-Ithar al-Marja'i al-Urubi al-Musytarak li al-Lughat li al-Thullab al-Jami'ah fi Indunisiya. *Arabiyatuna: Jurnal Bahasa Arab*, 7(1).
<https://doi.org/10.29240/jba.v7i1.6390>

Yul, W., Andrian, R., Musthofa, A., & Rozianie, J. F. (2022). Assessing Arabic speaking skills: A critical study for implementation of final examination in Indonesia. *Izdiyar: Journal of Arabic Language Teaching, Linguistics, and Literature*, 5(2), 151–166. <https://doi.org/10.22219/jiz.v5i2.22294>