



Arabic as an Additional Language: A Preliminary Scoping Synthesis of Pedagogy, Learner Processes, and Assessment

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 Thoriquattyas

Universitas Negeri Malang, Indonesia

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

Abstract

Arabic as a foreign or additional language is taught across markedly different learner populations, Arabic varieties, and instructional settings. This preliminary scoping synthesis examines how recent original studies frame pedagogy, learner processes, educational technologies, and assessment. Guided by the Population–Concept–Context framework and PRISMA-ScR, it charts 17 candidate studies identified through targeted publisher discovery; systematic database exports and independent screening remain outstanding. Within this candidate set, six studies address instruction, five digital resources, three learner processes, and three assessment or materials. Seven use an instructional comparison, whereas other studies primarily document experience, associations, or resources. Positive delayed performance findings coexist with null results, and reported engagement is not interchangeable with measured proficiency. The evidence-function taxonomy may support a future completed scoping review, but frequencies and research gaps cannot yet be generalized beyond this provisional corpus.

Keywords: Arabic language education; additional-language learning; evidence mapping; assessment

مستخلص البحث

يتناول هذا التحليل الاستكشافي الأولي البحوث المتعلقة بتعليم العربية للناطقين بغيرها وتعلّمها في سياقات تعليمية ولغوية متنوعة. وتهدف الدراسة إلى توصيف المناهج التدريسية وعمليات التعلّم والتقنيات التعليمية وأساليب التقويم، مع التمييز بين الانخراط في الأنشطة والتحصيل اللغوي المقاس. واسترشد العمل بإطار السكان والمفهوم والسياق وإرشادات التقارير للمراجعات الاستكشافية، واستند إلى خريطة أولية تضم سبع عشرة دراسة أصلية عُثر عليها عبر صفحات الناشرين. ولم تُستكمل بعد عمليات البحث المنهجية في قواعد البيانات أو فرز الدراسات بصورة مستقلة. وتناولت الدراسات المرشحة التدريس والتقنيات الرقمية ودافعية المتعلمين والقياس والمواد التعليمية، وتباينت في تصميماتها ومؤشرات نتائجها. وتظهر أمثلة على نتائج إيجابية وأخرى غير دالة إحصائية، مما يقتضي الحذر في تفسير فاعلية التدخلات. وتتلخص الدراسة إلى أن تصنيف الأدلة بحسب

موضوع القياس وقوة الاستدلال خطوة مفيدة، غير أن الاستنتاجات العامة تبقى مشروطة باستكمال إجراءات المراجعة وتوثيقها والتحقق من النصوص الكاملة.

الكلمات المفتاحية: تعليم العربية؛ اكتساب اللغة؛ رسم خريطة الأدلة؛ التقويم

1. Introduction

1.1. Context and significance

Arabic language education is not one pedagogical setting. Arabic may be taught as a foreign language in schools and universities outside Arabic-speaking communities, as an additional language within Arabic-speaking societies, or in curricula with pronounced religious, academic, professional, or intercultural objectives. The Arabic label also conceals distinctions between Modern Standard Arabic and spoken varieties. Such diversity has direct consequences for the transferability of teaching methods and the interpretation of outcomes.

1.2. Empirical research landscape

Recent primary research addresses instructional decisions at very different levels. For example, [Baalousha and Malovrh \(2023\)](#) contrasted guided induction and deductive instruction for Palestinian Arabic negation, whereas [Alzahrani and Almalki \(2024\)](#) experimentally examined narrative-based syntactic priming in Arabic L2 classrooms. On the learner side, [Calafato \(2023\)](#) followed the development of motivation and self-efficacy in two multilingual university learners, and [AlHroob and Al-khresheh \(2026\)](#) modeled associations between self-regulated learning strategies in a much larger university sample. These works contribute different kinds of evidence and therefore should not be pooled as if they answered one effectiveness question.

Technology constitutes another prominent strand. A Malaysian qualitative study explored tertiary students' experience of mobile games ([Abdul Ghani et al., 2022](#)); a quasi-experimental Indonesian study investigated text-to-speech support for listening and speaking ([Fajriah et al., 2025](#)); and a student survey explored interest in ChatGPT as a learning aid ([Murdani Z et al., 2025](#)). Their reported outcomes range from acceptability and engagement to assessed language performance. Conflating these categories risks overstating what technology integration has demonstrated.

1.3. Gap relative to previous reviews

Existing reviews have mapped gamification, technology-enhanced curricula, and Scopus-indexed Arabic-language research ([Almelhes, 2024](#); [Mohamed Sapawi & Nik Yusoff, 2025](#); [Zikriah & Mauludiyah, 2024](#)). These studies prevent a credible claim that the field is unreviewed. The narrower potential contribution here is an evidence-function framework that links each study to its Arabic variety, learners, setting, measured construct, comparison design, and permissible inference. Because the candidate set is not exhaustive, that contribution is presently demonstrated as a classification method, not as a definitive field-level gap analysis.

1.4. Objectives and contributions

Scoping methodology supports transparent mapping of concepts and evidence types rather than estimating a pooled intervention effect ([Munn et al., 2018](#)). The framework draws on JBI guidance and PRISMA-ScR reporting principles ([Pollock et al., 2024](#); [Tricco et al., 2018](#)). The provisional questions are: (RQ1) Which populations, settings, and Arabic varieties appear in the candidate literature? (RQ2) Which pedagogical and learner-related phenomena are examined, and with what designs? (RQ3) What do reported measures permit authors to conclude about engagement, attitudes, language performance, and retention?

1.5. Article organization

The following sections describe the preliminary eligibility and charting framework, map the candidate studies by their evidence objects, interpret the measurement and inference boundaries, and state the methodological work that remains necessary for a complete review.

2. Methods

2.1. Review design and status

The review protocol is informed by the Population–Concept–Context (PCC) framework. Population: non-native learners of Arabic and their teachers. Concept: pedagogy, learning activities, motivation, self-regulation, curriculum materials, digital tools, and measurement. Context: schools, universities, and organized courses across regions. Foreign, second, and additional language are retained as retrieval terms but must be coded as separate learning conditions when full texts disclose exposure to Arabic outside formal instruction. This is a protocol-informed preliminary synthesis; it is not a completed PRISMA-ScR review.

2.2. Eligibility criteria

The intended eligibility window is 1 January 2021 to 10 October 2026. Eligible evidence is original peer-reviewed research concerning non-native Arabic learners, their instruction, or purpose-built educational materials, including qualitative studies, experimental and quasi-experimental comparisons, psychometric studies, corpus development, and document analyses. Non-native learner status and Arabic variety must be identifiable or coded as 'not reported'. Reviews are contextual references, not primary evidence-map entries. Exclusively native-L1 literacy, unrelated Arabic NLP, duplicate datasets, and non-research commentary are excluded. Publication language alone is not a planned exclusion criterion where reliable translation and full-text review are possible.

2.3. Databases, Boolean search, and identification procedure

The pre-specified database plan covers Scopus, Web of Science Core Collection, ERIC, and supplementary Arabic-language and regional sources. Proposed English search: (Arabic OR "Arabic language" OR "Arabic as a foreign language" OR "Arabic as a second language") AND (teach* OR learn* OR pedagog* OR instruction* OR curriculum OR assess* OR motivation OR technolog*) AND ("non-native" OR "foreign language" OR "second language" OR learners OR students). An Arabic search component is ("تعليم اللغة العربية" OR "تعلم العربية" OR "العربية للناطقين بغيرها") AND (تعليم OR تعلم OR مهارات OR تقويم OR دافعية). Each platform requires tested field tags, exported complete results, language-specific query validation, and a recorded search date. These strings are proposed, not asserted to have been executed.

Targeted publisher-page discovery identified the 17 original reports provisionally charted here. It produced no auditable database-wide hit counts or deduplication file. Therefore, neither the 17 records nor the descriptive counts derived from them may be reported as a systematically selected sample. At full review stage, database exports, exact searches, dates, citation chasing, duplicate decisions, and exclusions must be archived.

2.4. Selection, charting, and synthesis

For the completed scoping review, titles and abstracts and then full texts require independent screening against the PCC criteria, reconciled decisions, and reasons for full-text exclusion. The pilot extraction schema records citation and DOI, country, learner status, age and proficiency, Arabic variety, setting, design, sample, comparator, exposure duration, language-skill outcome, instrument validity,

timing, and limitations. The candidate chart was not produced by two independent full-text reviewers; missing fields should not be interpreted as absent characteristics.

The present descriptive synthesis groups evidence objects and distinguishes implementation reports, cross-sectional associations, resource and instrument studies, and performance comparisons. No meta-analysis or causal ranking is attempted. Evidence-form counts concern only the 17 candidates and are descriptive inventory values, not prevalence estimates for Arabic-language research. Design-specific indicators, including baseline equivalence, follow-up, and instrument reliability, require full-text audit before a definitive chart can be published.

2.5. PRISMA-ScR flow and auditability

Figure 1 represents the intended PRISMA-ScR workflow. Search totals, duplicates, exclusions, retrieved reports, and final inclusion counts are deliberately unspecified because the systematic identification and screening process has not been completed. The charted 17 records must not be retrofitted into an invented flow diagram.

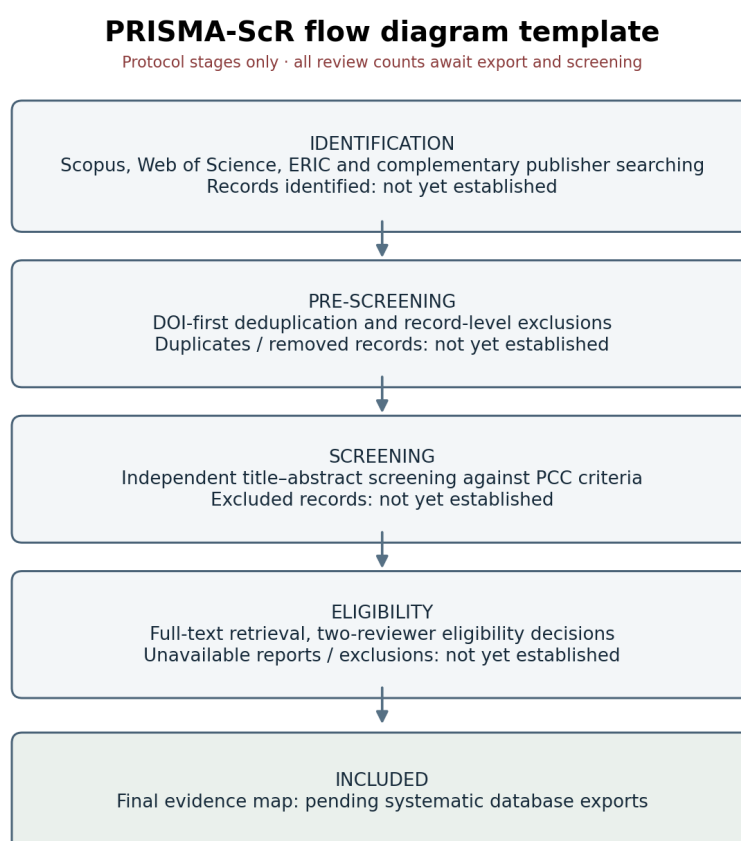


Figure 1. Proposed PRISMA-ScR workflow; numerical fields remain pending execution of systematic searching and screening.

3. Preliminary evidence map and results

The 17 candidate primary reports form a bounded developmental chart, not an estimate of the total number of eligible studies. Table 1 preserves the original study-by-study descriptions, whereas Table 2 classifies measurement and inferential boundaries. The counts presented in Table 3 are calculated directly from these candidate entries and cannot support conclusions about publication volume, global geographic concentration, or trends.

Table 1. Charted original studies: contexts, methods, and reported contributions

Study	Population / context	Design	Evidence object	Conservative evidence statement
S01 · Abdul Ghani et al. (2022)	Malaysia; tertiary Arabic classes	Qualitative interviews; n = 16 students	Mobile digital game-based vocabulary learning	Students described engagement, enjoyment, and perceived vocabulary benefits; no controlled effect estimate established by interviews.
S02 · Baalousha & Malovrh (2023)	AFL learners; spoken Palestinian Arabic	Instructional comparison; n = 27; delayed test	Guided induction versus deductive teaching of negation	Both groups gained immediately; guided induction outperformed deduction on a delayed test four weeks later.
S03 · Calafato (2023)	Norway; university Arabic	Longitudinal case study; n = 2; two semesters	Motivation, self-efficacy, multilingual repertoire	Motivation and self-efficacy varied over time; digital strategies did not uniformly preserve achievement.
S04 · Alshammari et al. (2024)	Saudi Arabia; learners of Arabic as L2	Multi-cohort instrument-development and validity study	Identity-related and learning-environment motivation	Developed and examined motivation subscales; instrument validity is the key contribution, not intervention effectiveness.
S05 · Alzahrani & Almalki (2024)	Arabic L2 classroom	Between-group experiment; n = 52; delayed posttest	Narrative-based syntactic priming	No significant increase in the target fronted-temporal structure at immediate or one-week follow-up.
S06 · Abu Issa (2024)	International school; grade 7 AFL	Quasi-experiment; n = 38; eight weeks	Project-based oral-language practice	Reported higher post-intervention speaking performance across assessed domains in the project group.

Study	Population / context	Design	Evidence object	Conservative evidence statement
S07 · Shofiyani (2025)	Indonesia; first-year university	Quasi-experimental non-equivalent groups	Think–Pair–Share for speaking	Publisher abstract reports improvement compared with conventional instruction.
S08 · Nashfati & Salamuddin (2025)	Indonesia; grade 7 madrasa	Qualitative classroom case study	Contextual teaching and learning	Observed active engagement and reported contextual comprehension; facility and time constraints remained.
S09 · Fajriah et al. (2025)	Indonesia; grade 11 madrasa	Quasi-experiment; experimental and control groups	Text-to-speech support for listening and speaking	Authors report gains, but post-speaking means reported as 85.72 versus 84.89; magnitude needs appropriate interpretation.
S10 · Rahma & Irma (2025)	Indonesia; grade 7 madrasa	Quasi-experiment; n = 40 (20/20)	Audio-lingual method	Compared speaking tests under audio-lingual versus routine instruction.
S11 · Shofiyani et al. (2025)	Indonesia; grade 10 madrasa	Quasi-experiment; n = 60; mixed data	Quizizz, leaderboard and interactive speaking tasks	Reported speaking improvement with d = 1.76 and higher participation; result requires full-text verification.
S12 · Murdani Z et al. (2025)	Indonesia; Arabic education undergraduates	Explanatory sequential mixed methods	Learner interest and use of ChatGPT	Describes interest and engagement determinants; does not independently establish learning gains.
S13 · Alhamami (2025)	India; university AFL learners	Cross-sectional models; n = 213	Attitudes, norms, control, enjoyment, anxiety and intention	Perceived behavioral control and subjective norms predicted intention; anxiety was negatively associated with intention.
S14 · Abazoğlu & Alhourani (2025)	Türkiye; Arabic–Turkish bilingual design	Corpus design and pilot mini-corpus; no classroom trial	Pedagogical corpus-based learning resources	Shows the feasibility and rationale for bilingual corpus tools, not comparative learner improvement.
S15 · Arkan (2026)	Türkiye; A1–A2 Arabic textbooks	Qualitative document analysis	CEFR alignment of listening activities	Identified 118 A1 and 46 A2 listening activities; noted shortcomings in pre-listening and multimedia provision.
S16 · AlHroob & Al-khresheh (2026)	Jordan; university Arabic programs	Cross-sectional SEM; n = 318	Self-monitoring, goal-setting, emotion regulation	Reported associations and moderated links among strategies; no experimental causal inference.

Study	Population / context	Design	Evidence object	Conservative evidence statement
S17 · Fahmi (2026)	United States-focused AFL textbook series	Critical content analysis; five textbook volumes	Cultural representation in Al-Kitaab and Arabiyyat al-Naas	Found an emphasis on surface/objective cultural framing relative to subjective/deep culture.

Table 1 distinguishes learners from textbooks and other educational artifacts as units of analysis. A student's description of an enjoyable game, a questionnaire score for learning intention, and a delayed grammar-production assessment each address different outcomes. This distinction prevents unmeasured proficiency claims from being attributed to descriptive studies.

Figure 2 makes the logic of the proposed extraction framework explicit: populations and contexts shape exposure to teaching strategies and learner processes; studies then record different outcomes through instruments with varying inferential strength. The diagram is a charting model, not a proposed causal mediation model. It makes clear why interpretation must track age group, Arabic variety, proficiency, measurement, and the presence or absence of a comparator.

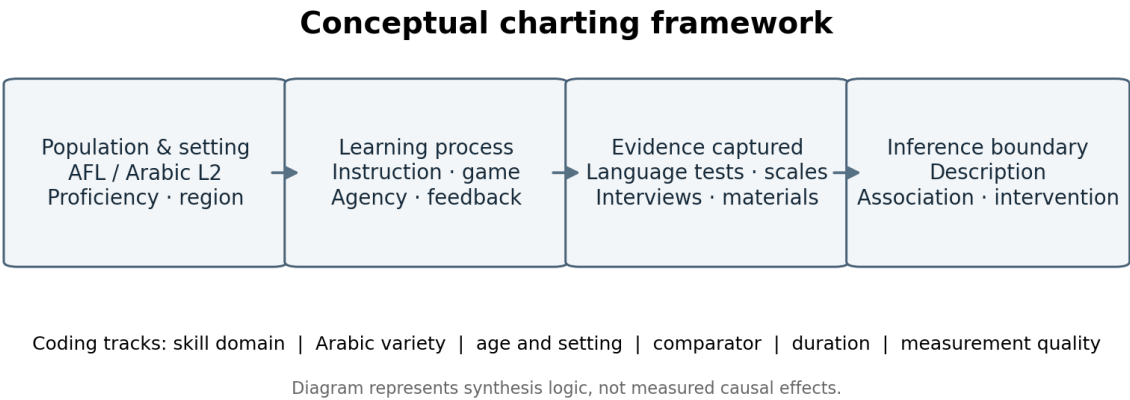


Figure 2. Conceptual charting framework for Arabic-language education research (author-created schema).

Table 2 exposes differences among self-report, intervention comparison, documentary content coding, and psychometric validity evidence. A digital technology can be feasibly implemented without demonstrating greater retained proficiency. The separate outcome-type categories in Table 3 make the descriptive imbalance in this candidate set visible without treating 'comparative' as synonymous with causally valid.

Table 2. Study-by-study measurement, limitations, and evidence classification

Study	Outcomes and measurement	Interpretive limitation	Mapping strand	Evidence form
S01 · Abdul Ghani et al. (2022)	Semi-structured interviews and perceptions	Self-report and small purposive group; avoid calling improvement causal.	Digital	Non-comparative
S02 · Baalousha & Malovrh (2023)	Accuracy and response times in pretest/posttest/delayed task	Small sample and a specific linguistic structure.	Instruction	Comparative
S03 · Calafato (2023)	Weekly interviews, rating logs, exam scores	Two-person longitudinal study offers depth, not population estimates.	Learner	Non-comparative
S04 · Alshammari et al. (2024)	Survey subscales, reliability and validation interviews	Some cohorts could overlap; transportability to other regions requires checking.	Measurement	Non-comparative
S05 · Alzahrani & Almalki (2024)	Structure-use baseline, immediate and delayed tasks	Null finding concerns a specific syntactic form and implementation.	Instruction	Comparative
S06 · Abu Issa (2024)	Observation checklist; 27 subskills	Small, non-randomized cohort; journal and assessment details need deeper audit.	Instruction	Comparative
S07 · Shofiyani (2025)	Pre/post speaking tests and classroom observation	Year/volume metadata differs across hosted copies; cite online 2025 and verify production record.	Instruction	Comparative
S08 · Nashfati & Salamuddin (2025)	Observation, interviews, documents	Single classroom; no independent proficiency effect estimation.	Instruction	Non-comparative
S09 · Fajriah et al. (2025)	Listening and speaking pre/post scores	Small between-group post-score contrast; baseline and statistics must be checked.	Digital	Comparative
S10 · Rahma & Irma (2025)	Pre/post speaking performance and analytic checklist	Intact classes; generalizability and baseline equivalence require review.	Instruction	Comparative

Study	Outcomes and measurement	Interpretive limitation	Mapping strand	Evidence form
S11 · Shofiyani et al. (2025)	Speaking tests, questionnaires, observation, interviews	Single-site, possible novelty effects and outcome measurement bias.	Digital	Comparative
S12 · Murdani Z et al. (2025)	Survey and qualitative follow-up	Interest and perceived utility are not demonstrated proficiency.	Digital	Non-comparative
S13 · Alhamami (2025)	Psychological survey measures and SEM modelling	Associational intention models cannot establish sustained language performance.	Learner	Non-comparative
S14 · Abazoğlu & Alhourani (2025)	Corpus construction and illustrative analysis	No intervention, retention, or proficiency evidence.	Digital	Non-comparative
S15 · Arkan (2026)	Textbook activity coding against CEFR descriptors	One selected teaching set; activity counts are not learning gains.	Measurement	Non-comparative
S16 · AlHroob & Al-khresheh (2026)	Validated self-regulation questionnaire	Cross-sectional associations and self-report.	Learner	Non-comparative
S17 · Fahmi (2026)	Deductive and inductive coding of five volumes	Textbook discourse analysis cannot measure students' intercultural competence.	Measurement	Non-comparative

Note. Classification refers to the study design as summarized in available published records, not independently appraised risk of bias. Study details, including missing or delayed assessments, are subject to full-text confirmation.

Table 3. Provisional evidence strands among 17 candidate studies

Evidence strand	Comparative	Non-comparative	Total candidates
Instruction	5	1	6
Digital	2	3	5
Learner	0	3	3
Measurement/ materials	0	3	3
Total	7	10	17

Note. Candidate-chart counts only; no systematic search denominator exists. “Comparative” identifies a reported group contrast rather than established causal validity.

The distinction in Table 3 is cross-checked against both Table 1 (design and context) and Table 2 (measurement and inference). The seven comparisons concern instructional or digital interventions; the learner-process and material-centered strands are mainly observational, psychometric, or documentary. Figure 2 visually explains why these strands require different charting variables.

3.1. Instructional approaches and language performance

The instructional strand includes guided induction, narrative priming, project-based learning, cooperative discussion, contextual teaching, and structured oral drills. Importantly, these designs do not all converge on significant language gains. Baalousha and Malovrh (2023) observed comparatively durable accuracy after guided-inductive instruction on a specific Palestinian Arabic negation form. In contrast, Alzahrani and Almalki (2024) reported no significant increase in the target temporal-phrase structure after narrative priming, including at delayed posttest. Taken together, these cases suggest that exposure, the linguistic target, the learning task, and the temporal horizon are integral to any claim about pedagogical transfer.

Quasi-experimental studies in school and university settings add different evidence. Abu Issa (2024) reported speaking gains after an eight-week project-based sequence, Shofiyani (2025) investigated Think–Pair–Share in an Indonesian university, and Rahma and Irma (2025) compared audio-lingual instruction with routine teaching in intact classes. While such studies are directly relevant to classroom decision-making, small samples, non-random group allocation, and site-specific norms may distort generalization. Nashfati and Salamuddin (2025) instead described contextual pedagogy in a naturalistic grade-seven setting; classroom participation

should not be transformed into an experimentally identified improvement in proficiency.

3.2. Technology, engagement, and attribution

Digital games and language technologies appear in sharply different forms. Abdul Ghani et al. (2022) interviewed 16 tertiary students about mobile games and identified positive learner experiences, but interview testimony is not equivalent to controlled vocabulary gain. Shofiyani et al. (2025) reported a large effect on assessed speaking alongside observations and questionnaires, although independent replication and measurement scrutiny remain warranted. Fajriah et al. (2025) reported gains with text-to-speech media, while the published abstract's small difference in posttest speaking means calls for closer inspection of baseline adjustment and uncertainty, rather than an unqualified endorsement of the tool.

Generative AI creates a similar inferential trap. The study by Murdani Z et al. (2025) is explicitly about students' interest and engagement with ChatGPT, not a causal comparison of language attainment. Abazoğlu and Alhourani (2025) offer a bilingual corpus design for Turkish learners, providing a practical resource model without testing a classroom intervention. These examples justify classifying studies by the question their design can actually answer: feasibility, uptake, teacher enactment, skill performance, or durability of learning.

3.3. Learner motivation and self-regulation

Learner-process studies enrich the map but must not be mistaken for intervention studies. Calafato (2023) traced within-person fluctuations in motivation and self-efficacy in two multilingual Norwegian university students; the detailed temporal account challenges simplistic assertions that motivation rises monotonically with educational support. Alshammari et al. (2024) addressed the construction and validation of Arabic-L2 motivation measures, making measurement itself the object of study. Alhamami (2025) found associations between intention to learn AFL, perceived behavioral control, social norms, anxiety, and other psychosocial constructs in an Indian university sample; the cross-sectional design supports prediction of intention, not a causal explanation of proficiency.

AlHroob and Al-khresheh (2026) modeled links among self-regulated learning strategies in 318 non-native Arabic learners in Jordan. Reported moderation by proficiency and learning duration suggests that a strategy's apparent relevance may differ across learner trajectories. Across these studies, stable definitions and validated Arabic-specific instruments are needed, but they must be aligned with

performance outcomes if the field seeks to connect motivation, autonomy, and language competence.

3.4. Curriculum and materials as evidence objects

The curriculum-material studies underscore the difference between analyzing educational inputs and measuring learner attainment. [Arkan \(2026\)](#) coded 118 A1 and 46 A2 listening activities in one Arabic textbook series and discussed alignment with CEFR descriptors; these are properties of the teaching materials, not student test results. [Fahmi \(2026\)](#) examined cultural representations across five textbook volumes using a critical content approach. Neither study can independently establish improvements in listening comprehension or intercultural competence.

4. Discussion

4.1. An evidence map, not a league table of pedagogies

The core analytical contribution is to classify evidence by its object and permissible inference. Among the 17 candidates, the chart distinguishes six instruction-oriented reports, five digital resource reports, three learner-process reports, and three assessment/material studies (Table 3). Seven candidates contain an instructional comparison and ten do not. These figures are useful for diagnosing the current chart and checking coding coherence, but they are not evidence of the field's true disciplinary distribution. Tables 1–3 show why a consolidated 'most effective method' ranking would collapse non-commensurable designs and outcomes.

A particularly important contrast is the tension between positive and null results. The guided-induction study documents sustained advantages on a defined grammatical target ([Baalousha & Malovrh, 2023](#)), while the narrative priming study did not find the hypothesized structure-use change ([Alzahrani & Almalki, 2024](#)). There is no methodological basis for privileging positive outcomes solely because they translate more readily into classroom recommendations. Differences in treatment duration, instruction intensity, target structures, linguistic varieties, and test alignment may explain variation, but those explanations remain hypotheses requiring direct comparative evidence.

4.2. Beyond the technology-versus-tradition contrast

The technologically focused papers invite more precise questions than whether digital Arabic teaching 'works'. In practice, games may change opportunities for retrieval and collaborative interaction; text-to-speech may increase access to modeled pronunciation and listening input; chatbots may provide rapid feedback

or simply become attractive tools; and learner corpora may expand authentic example access. These are different pedagogical mechanisms. To isolate them, future studies should document the interaction task, amount of exposure, teacher feedback, linguistic variety, and actual time on task rather than merely label one arm as 'technology' ([Abdul Ghani et al., 2022](#); [Fajriah et al., 2025](#); [Abazoğlu & Alhourani, 2025](#)).

Acceptability is an important implementation outcome but cannot stand in for attainment. Students may enjoy mobile games and be interested in ChatGPT without demonstrating delayed transfer to independent speaking or writing. For that reason, observations and interviews should be paired with valid language assessments and, where possible, pre-registered comparison designs. Conversely, null effects in a task-specific study are informative, especially when delayed tests reveal whether learning survives beyond the instructional session.

4.3. Arabic-specific linguistic and cultural constraints

The Arabic-learning context cannot be reduced to the generic English-as-a-foreign-language template. The curricular target may be Modern Standard Arabic, a spoken variety, or a staged relationship between both. [Baalousha and Malovrh \(2023\)](#) make the variety explicit by investigating spoken Palestinian Arabic. [Arkan \(2026\)](#) demonstrates that listening activity design can be examined against CEFR categories without assuming identical linguistic realization across textbooks and varieties. Future synthesis must chart script familiarity, diglossic exposure, home languages, proficiency standards, cultural orientation, and language-use setting rather than adding these variables as afterthoughts.

[Fahmi's \(2026\)](#) textbook analysis raises an additional concern: teaching materials may reproduce narrow cultural representations even when lessons are communicatively well designed. Cultural content is not a decorative supplement to language competence; it shapes whose Arabic is normalized and which communities become visible. However, critical analysis of textual representation is still distinct from a measured effect on intercultural awareness. A rigorous evidence map should keep these two layers visible, while encouraging studies that connect materials analysis to learner interpretation.

4.4. Rebuilding outcome measurement

The candidate evidence exposes three distinct measurement problems: construct substitution, timing, and cross-context comparability. Reported interest or perceived usefulness measures acceptability, not speaking proficiency; textbook

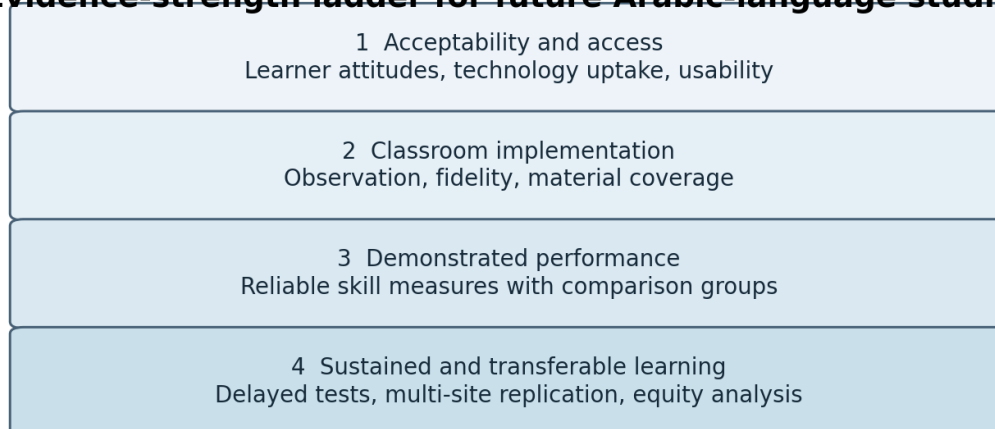
frequencies describe resources, not comprehension; and a single immediate posttest may not indicate retained language learning. Assessment claims therefore depend on specified linguistic skill, test validity, rating procedure, comparator, and follow-up interval. Table 2 records the available measurement type, while Figure 2 links it to learner context and inference boundaries.

Alshammari et al. (2024) make a constructive contribution by explicitly testing motivational instruments for Arabic L2 learners, whereas Arkan (2026) supplies a separate model of materials coding with articulated descriptors. Their common lesson is not that all studies need the same instrument; rather, each type of claim requires measures designed for the construct at issue. The proposed field extraction in Figure 2 therefore links evidence object and measure before any synthesis of purported effects.

4.5. A tiered future-research agenda

Figure 3 is an author-generated heuristic about types of evidence, not an observed hierarchy of research quality or a PRISMA outcome. Studies of usability and classroom implementation can be rigorous for their own descriptive questions; intervention effects require appropriately measured outcomes and credible comparisons. The diagram is used to distinguish claims rather than to devalue interpretive or qualitative research.

Evidence-strength ladder for future Arabic-language studies



Higher rungs support stronger claims; none is guaranteed by a digital tool alone.

Figure 3. Evidence-strength ladder for Arabic teaching research. Higher rungs demand different inference, not a blanket judgment of study value.

The hierarchy in Figure 3 also creates a way to organize the next generation of studies without imposing one pedagogy across all Arabic contexts. For a mobile-

game intervention, investigators might first examine student access and acceptability, then implementation fidelity, then speaking or vocabulary tests, and finally transfer to unguided production after several weeks. For a corpus tool, the analogous path starts with usability of examples and learners' search behavior before moving to assessed inference about morphosyntax or lexical choice. The proposed pathway places research design ahead of promotional technology labels.

4.6. The specific gap this scoping review could fill

Existing reviews already synthesize gamification and digital Arabic curricula ([Almelhes, 2024](#); [Mohamed Sapawi & Nik Yusoff, 2025](#)). The proposed incremental advance lies in classifying heterogeneous research by target language variety, educational setting, directness of learning measurement, comparator and follow-up. This candidate-level demonstration suggests how a completed evidence map could distinguish topics ripe for a focused effectiveness review from topics requiring concept development or new measurement tools.

5. Limitations and planned validation

Major methodological limitations remain. The identified records did not arise from completed Scopus, Web of Science, ERIC, or Arabic-language index exports, and full-text independent selection, deduplication, and PRISMA reconciliation have not occurred. Consequently, the study cannot support population-level claims about regional coverage, frequency of research designs, or the prevalence of methodological weaknesses. Any observed asymmetry in this candidate set may be a feature of targeted retrieval rather than the scholarly literature.

The first author-metadata anomaly requiring reconciliation concerns Shofiyani's Think–Pair–Share study: the journal landing page places the article with a 2025 issue/DOI record, while the hosted PDF identifies a 2024 volume/issue. A formal evidence map would also require verified language backgrounds, Arabic variety, baseline equivalence, follow-up measures, and the reported scores for each full text. Until these checks are complete, conclusions must remain provisional.

6. Conclusion

This preliminary chart shows how original Arabic additional-language research can be differentiated by what it observes: teaching implementation, learner perceptions, psychosocial processes, performance comparisons, psychometric instruments, or instructional materials. In the candidate set, positive and null findings coexist, while measures of engagement cannot be taken as substitutes for measured language

attainment. The chief contribution is a transparent evidence-function taxonomy, not an exhaustive map or an effectiveness estimate. A publishable PRISMA-ScR review depends on reproducible multilingual searching, independent study selection, reconciled citation metadata, and full-text verification.

References

- Abazoğlu, M., & Alhourani, M. I. (2025). The use of language corpora in teaching Arabic to Turkish speakers within the framework of computational linguistics. *Social Sciences & Humanities Open*, 12, 101947. <https://doi.org/10.1016/j.ssaho.2025.101947>
- Abdul Ghani, M. T., Hamzah, M., Wan Daud, W. A. A., & Muhamad Romli, T. R. (2022). The impact of mobile digital game in learning Arabic language at tertiary level. *Contemporary Educational Technology*, 14(1), ep344. <https://doi.org/10.30935/cedtech/11480>
- Abu Issa, H. H. I. (2024). The effect of project-based learning on developing speaking skills among Arabic as a foreign language learners in international schools: A multidimensional quasi-experimental study. *SCHOLAR*, 4(1), 1–43. <https://doi.org/10.66773/f0tajp21>
- Alhamami, M. (2025). Intention over motivation: A holistic analysis of psychological constructs in Arabic as a foreign language learning. *Acta Psychologica*, 258, 105142. <https://doi.org/10.1016/j.actpsy.2025.105142>
- AlHroob, B. F., & Al-khresheh, M. H. (2026). Self-regulated learning strategies among non-native Arabic learners: A structural equation modelling approach. *Asian-Pacific Journal of Second and Foreign Language Education*, 11, Article 8. <https://doi.org/10.1186/s40862-025-00380-7>
- Almelhes, S. A. (2024). Gamification for teaching the Arabic language to non-native speakers: A systematic literature review. *Frontiers in Education*, 9, 1371955. <https://doi.org/10.3389/feduc.2024.1371955>
- Alshammari, A., Shaalan, D., & Al-onazi, B. (2024). Instruments measuring motivation to learn Arabic as a second language: evidence of validity and reliability. *Humanities and Social Sciences Communications*, 11, Article 887. <https://doi.org/10.1057/s41599-024-03381-y>
- Alzahrani, A., & Almalki, H. (2024). Syntactic priming in the classroom: Using narratives to prime L2 Arabic speakers. *Asian-Pacific Journal of Second and Foreign Language Education*, 9, Article 68. <https://doi.org/10.1186/s40862-024-00292-y>
- Arkan, Z. (2026). Evaluation of listening activities in Arabic language teaching textbooks used in Turkey in light of the CEFR criteria. *Asian-Pacific Journal of Second and Foreign Language Education*, 11, Article 10. <https://doi.org/10.1186/s40862-025-00366-5>

- Baalousha, H., & Malovrh, P. A. (2023). Guided induction and deductive instruction: Analyzing efficiency versus depth of processing in the L2 acquisition of Palestinian Arabic. *Ampersand*, 11, 100140. <https://doi.org/10.1016/j.amper.2023.100140>
- Calafato, R. (2023). Charting the motivation, self-efficacy beliefs, language learning strategies, and achievement of multilingual university students learning Arabic as a foreign language. *Asian-Pacific Journal of Second and Foreign Language Education*, 8, Article 20. <https://doi.org/10.1186/s40862-023-00194-5>
- Fahmi, Z. (2026). The cultural politics of curriculum discourse: Cultural representations in Arabic as a foreign language (AFL) textbooks. *Linguistics and Education*, 92, 101518. <https://doi.org/10.1016/j.linged.2026.101518>
- Fajriah, A., Mutia Tsani, N., Mulyaningsih, D., & Hasugian, K. N. (2025). Using text-to-speech media to improve istima' and kalam skills in Arabic language learning. *Tadris Al-'Arabiyyah*, 4(1), 1–14. <https://doi.org/10.15575/ta.v4i1.44854>
- Mohamed Sapawi, M. S., & Nik Yusoff, N. M. R. (2025). Integrating technology into the Arabic language curriculum: A systematic review of trends, strategies and cultural dimensions. *Social Sciences & Humanities Open*, 12, 101974. <https://doi.org/10.1016/j.ssaho.2025.101974>
- 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, Article 143. <https://doi.org/10.1186/s12874-018-0611-x>
- Murdani Z, E., Rahman, N. F., Yasiron, A. M., Roziqi, M. A., & Ilmiani, A. M. (2025). Minat mahasiswa menggunakan ChatGPT dalam pembelajaran bahasa Arab: Studi pada mahasiswa PBA UINSI Samarinda. *Borneo Journal of Language and Education*, 5(1), 155–168. <https://doi.org/10.21093/benjole.v5i1.10189>
- Nashfati, N., & Salamuddin (2025). Contextual teaching and learning in Arabic as a foreign language: A classroom-based case study in Indonesia. *Alibbaa': Jurnal Pendidikan Bahasa Arab*, 6(2), 356–372. <https://doi.org/10.19105/ajpba.v6i2.21278>
- Pollock, D., Peters, M. D. J., Tricco, A. C., Munn, Z., Jia, R. M., Alexander, L., Pieper, D., Evans, C., Godfrey, C. M., Brandão de Moraes, É., Saran, A., Campbell, F., & Khalil, H. (2024). Scoping reviews. In *JBIM Manual for Evidence Synthesis* (2024 ed.). JBI. <https://doi.org/10.46658/JBIMES-24-09>
- Rahma, Y. A., & Irma, A. (2025). Drilling toward fluency: Effects of the audio-lingual method on grade-VII Arabic speaking at state Islamic junior high school. *Lingua: Jurnal Keilmuan dan Kependidikan Bahasa Arab*, 11(1), 20–30. <https://doi.org/10.32678/lingua.v11i1.12102>

- Shofiyani, A. (2025). Model pembelajaran Think Pair Share dalam meningkatkan kemampuan berbicara bahasa Arab di perguruan tinggi. *Aphorisme: Journal of Arabic Language, Literature, and Education*, 5(2), 134–154. <https://doi.org/10.37680/aphorisme.v5i2.6568>
- Shofiyani, A., Mahliatussikah, H., & Barry, M. Y. F. (2025). Efektivitas gamifikasi dalam meningkatkan keterampilan berbicara Arab siswa Madrasah Aliyah. *Aphorisme: Journal of Arabic Language, Literature, and Education*, 6(2). <https://doi.org/10.37680/aphorisme.v6i2.8692>
- 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>
- Zikriah, Z., & Mauludiyah, L. (2024). Research trend of Arabic language teaching in the world: Systematic literature review based on Scopus database. *Research and Development in Education (RaDEn)*, 4(1), 27–48. <https://doi.org/10.22219/raden.v4i1.31847>

