Journal of Insights in Basic and Applied Sciences
مجلة رؤى في العلوم الأساسية والتطبيقية
ISSN: 3106-1737
Volume 2, Issue 4, 2026, Pages: 552-585
Journal homepage: https://jibas.com.ly/index.php/jibas/en

Listening Skill Challenges in Libyan EFL Higher Education: A Systematic Review of Empirical and Contextual Evidence (2015–2026)

Asmaa Ramadhan Albilaezi[1 *], Samera Salem Alzwam[2],
[1,2] Department of English, Faculty of Education, Elmergib University, Libya

تحديات مهارة الاستماع في سياق تعليم اللغة الإنجليزية كلغة أجنبية (EFL) في مؤسسات التعليم العالي بليبيا: مراجعة منهجية للأدلة التجريبية والسياقية(2015–2026)

أسماء رمضان البليزي[1*]، سامرة سالم الزوام[2]
[2،1] قسم اللغة الإنجليزية، كلية التربية، جامعة المرقب، ليبيا

*Corresponding author: aralbilaezi@elmergib.edu.ly

Received: September 26, 2026	Accepted: October 25, 2026	Published: November 26, 2026

Copyright: © 2026 by the authors. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract:
Listening comprehension remains a significant challenge in English language education, yet evidence concerning listening difficulties in Libyan EFL higher education is fragmented and insufficiently synthesised. This systematic review aims to identify, appraise, and synthesise empirical and contextual evidence on these challenges between 2015 and 2026. A systematic search was conducted across ERIC, Crossref, DOAJ, Dimensions, Google Scholar, ResearchGate, and Libyan university repositories and journal platforms. The review followed PRISMA 2020 guidelines; methodological quality was assessed using MMAT (2018), while certainty of evidence was evaluated using GRADE. Given the heterogeneity of the evidence, findings were synthesised narratively and thematically. Of 346 records identified, 15 reports representing 14 studies and 12 independent samples were included. These comprised six direct university-level listening samples, one teaching-and-assessment study, six contextual sources, and two intervention studies. Recurring challenges involved connected speech, lexical processing, cognitive and metacognitive demands, and affective factors, alongside pedagogical, assessment-related, and environmental constraints. Certainty was low for perceptual findings and very low for the remaining evidence. This review makes three contributions: (1) Evidence Mapping, by systematically mapping available empirical and contextual evidence; (2) Evidence Differentiation, by distinguishing direct university-level evidence from teaching-and-assessment, contextual, and intervention evidence; and (3) Research Gap Identification, by highlighting methodological limitations, particularly the scarcity of objective listening-performance measures and rigorous university-level intervention studies. These contributions provide an evidence-informed foundation for future research, instructional improvement, and educational decision-making in Libyan higher education.
.

Keywords: listening comprehension; listening difficulties; Libyan EFL learners; higher education; systematic review; listening instruction; methodological quality.

الملخص:
تظل مهارة الاستيعاب السمعي تحدياً جوهرياً في تعليم اللغة الإنجليزية، إلا أن الأدلة المتعلقة بصعوبات الاستماع في سياق التعليم العالي في ليبيا - حيث تُدرَّس الإنجليزية كلغة أجنبية- لا تزال مشتتة وتفتقر إلى التوليف الشامل. تهدف هذه المراجعة المنهجية إلى تحديد وتقييم وتوليف الأدلة التجريبية والسياقية حول هذه التحديات خلال الفترة الممتدة من عام 2015 إلى عام 2026. أُجري بحث منهجي عبر قواعد بيانات ومصادر متنوعة، شملت : ERIC، وCrossref، وDOAJ، وDimensions، وGoogle Scholar، وResearchGate، بالإضافة إلى مستودعات الجامعات الليبية ومنصات المجلات العلمية فيها. اتبعت المراجعة إرشادات"PRISMA 2020" ؛ حيث تم تقييم الجودة المنهجية باستخدام أداة "MMAT" (إصدار 2018)، بينما قُيِّمت درجة اليقين في الأدلة باستخدام نظام ."GRADE" ونظراً لتنوع الأدلة وتباينها، جرى عرض النتائج وتوليفها بأسلوب وصفي

وموضوعي. ومن بين 346 سجلاً تم تحديدها، أُدرجت 15 وثيقة، تمثل 14 دراسة و12 عينة مستقلة. وتضمنت هذه الوثائق: ست عينات مباشرة من مستوى التعليم الجامعي، ودراسة واحدة حول التدريس والتقييم، وستة مصادر سياقية، ودراستين حول التدخلات التعليمية. وقد تمثلت التحديات المتكررة في: خصائص الكلام المتصل، ومعالجة المفردات، والمتطلبات المعرفية وما وراء المعرفية، والعوامل الوجدانية، إلى جانب قيود تتعلق بطرق التدريس، والتقييم، والبيئة التعليمية. وقد وُجدت درجة يقين منخفضة فيما يخص النتائج المتعلقة بالإدراك السمعي، بينما كانت درجة اليقين منخفضة جداً بالنسبة لبقية الأدلة. تقدم هذه المراجعة ثلاث إسهامات رئيسية: (1) رسم خرائط الأدلة، من خلال تصنيف الأدلة التجريبية والسياقية المتاحة بشكل منهجي؛ (2) التمييز بين الأدلة، عبر الفصل بين الأدلة المباشرة من مستوى التعليم الجامعي وبين الأدلة المتعلقة بالتدريس والتقييم، والسياق العام، والتدخلات التعليمية؛ و(3) تحديد الفجوات البحثية، من خلال تسليط الضوء على أوجه القصور المنهجية، لا سيما ندرة المقاييس الموضوعية لأداء الاستماع، وغياب الدراسات التدخلية الرصينة في الجامعات. وتوفر هذه الإسهامات أساساً يستند إلى الأدلة لدعم البحوث المستقبلية، وتطوير الممارسات التدريسية، واتخاذ القرارات التعليمية في قطاع التعليم العالي في ليبيا.

الكلمات المفتاحية: الاستيعاب السمعي؛ صعوبات الاستماع؛ متعلمو اللغة الإنجليزية كلغة أجنبية في ليبيا؛ التعليم العالي؛ مراجعة منهجية؛ تدريس مهارة الاستماع؛ الجودة المنهجية.

1. Introduction

1.1 Background

Listening is the primary channel of comprehensible input through which vocabulary, phonology, and syntax are internalised, and it has long been described as the most problematic skill for foreign language learners (Omar, 2022, p. 6). Two distinctions frame this review: listening as a curricular skill differs from listening comprehension as a cognitive process, and perceived difficulty differs from measured proficiency; most available Libyan evidence is perceptual, a limit carried through every conclusion.

In Libya, English is taught as a foreign language throughout general education and at university level, where English departments offer dedicated listening courses. The national evidence is consistent with three conditions, none established nationally: university students and teachers report that listening courses are under-implemented and often reduced to testing (Al-shareef, 2022; Mohammad & Alsharf, 2025; Zenbil, 2021); school-level studies document preparation conditions preceding university entry, including exclusion of listening from examinations (Alghanimi & Ghazalla, 2021; Pathan et al., 2016); and recent review and conceptual sources reach compatible conclusions (Bahron & Abujela, 2026; Bin Naji et al., 2026). Only the first group is direct university evidence; the others are contextual (Section 5.6).

1.2 Statement of the Problem

The problem has four elements. Studies of listening in Libyan higher education are dispersed across national journals, repositories, and regional outlets, and no published systematic synthesis of them was identified in the searches conducted, a claim bounded by the search restrictions disclosed in Section 8. The studies vary widely in design, samples, and instruments, so their findings cannot be read as interchangeable. No previous account has distinguished measured from perceived difficulty, or direct university from school-level evidence, or rated the certainty of any conclusion.

1.3 Research Questions

1. What are the methodological, demographic, and institutional characteristics of the studies that investigated listening difficulties among English language learners in Libyan higher education between 2015 and 2026?
2. What linguistic, perceptual, cognitive, affective, pedagogical, and environmental challenges are reported in the included studies, and to what extent are these challenges supported by direct, contextual, or intervention evidence?
3. What do the studies reveal about how the listening skill is taught and assessed in Libyan higher education?
4. What educational interventions have been tested, what are the limits of their results, and what research gaps does the available evidence reveal?

1.4 Objectives

The review (a) maps the national evidence base on listening challenges in Libyan higher education; (b) classifies each source as direct, teaching-assessment, contextual, or intervention evidence; (c) appraises methodological quality and risk of bias with standardised instruments; (d) synthesises the challenges into a thematic framework with explicit certainty ratings; and (e) formulates a tiered agenda for research, practice, and policy.

1.5 Significance

The review offers a PRISMA-reported, criterion-appraised baseline for a national literature effectively absent from indexed international outlets. It reports what the literature shows, at what certainty, and where the evidence stops, rather than estimating national prevalence. As a result, this study contributes to an evidence-informed foundation for future research, instructional improvement, and educational decision-making in Libyan higher education.

2. Theoretical Framework

2.1 Cognitive Models of Listening Comprehension

The reviewed literature is dominated by Anderson's (1995) three-phase model: perception (encoding acoustic input and maintaining attention), parsing (converting input into a mental representation of meaning), and

utilisation (relating that representation to context and intention) (pp. 378-379). The phases are recursive rather than strictly sequential (Goh, 2000), and the model is diagnostic rather than explanatory: it classifies loci of difficulty rather than measuring performance uniformly. Omar (2022) adopted it as the outcome framework for Libyan learners' reported processing problems, measured through a perceived-problems questionnaire rather than direct measurement of each phase.

2.2 Bottom-Up and Top-Down Processing

Listening depends on the interaction of bottom-up decoding (extracting sounds, words, and structures from the signal) and top-down interpretation (using context and background knowledge). Weak bottom-up decoding forces compensatory top-down guessing, which breaks down where input is lexically dense or unfamiliar. Field (2008) argues that this calls for instruction that trains the decoding of natural speech rather than leaving comprehension to general guessing, a position the intervention evidence bears out indirectly (Section 5.7).

2.3 The Metacognitive Perspective

The metacognitive perspective treats listening as a strategic, self-regulated process taught through planning, monitoring, evaluation, and problem solving (Vandergrift & Goh, 2012). It is operationalised in the corpus by the Metacognitive Awareness Listening Questionnaire (MALQ) administered at Omar Al-Mukhtar University (Bianku & Mahmud, 2026). One caution governs its use: MALQ measures awareness of strategies, not listening proficiency.

2.4 A Classificatory Taxonomy of Difficulty Sources

For classifying difficulty sources, the review uses the four-factor taxonomy reported in the national literature: "listening difficulties have four primary causes: the speaker, the listener, the message content, and the physical surrounding" (Yagang, 1994, as cited in Othman, 2025, p. 177). The taxonomy is a classificatory device for organising the evidence, not a tested causal theory.

2.5 Theoretical Integration and Analytical Framework

The strands combine into one analytical framework used for extraction and synthesis: acoustic and instructional inputs feed perceptual processing; perceptual output feeds parsing, which feeds utilisation; working memory and attention constrain all three phases; lexical knowledge supplies the units parsing operates on; metacognitive regulation and affective state moderate the sequence; and an institutional layer (curricula, teaching practices, assessment policy, teacher education, infrastructure) shapes every input and regulator. Table 1 states each component's function. The framework is an analytical device, not a validated causal model; supported and hypothetical links appear in Section 5.9.

Table 1. The analytical framework: components, functions, and exemplifying evidence.

Component	Analytical function	Exemplifying evidence
Acoustic and instructional inputs	Classify signal and teaching conditions	Speech rate, connected speech, accents, audio quality (Omar, 2022; Mohammad & Alsharf, 2025)
Perceptual processing	Locate difficulties in recognising sounds and words	Speed and connected-speech items; test-based recognition deficits (Omar, 2022; Zenbil, 2021)
Cognitive processing	Classify memory, attention, and inference demands	Parsing means 3.63-3.79; real-time inference items (Omar, 2022; Othman, 2025)
Affective factors	Describe anxiety and confidence as reported	Anxiety item M = 3.94; 48.2% fear not understanding (Omar, 2022; Othman, 2025)
Metacognitive regulation	Describe planning, monitoring, evaluation	MALQ profile (Bianku & Mahmud, 2026)
Environment and institution	Organise teaching, assessment, resource evidence	Course non-delivery, washback, equipment (Othman, 2025; Alghanimi & Ghazalla, 2021; Al-shareef, 2022)

3. Review Design and Protocol

3.1 Design Statement

This study employed a systematic review design to identify, critically appraise, and synthesise empirical and relevant contextual evidence on listening skill challenges in Libyan higher education, following PRISMA 2020 (Page et al., 2021). Given the heterogeneity of designs, populations, instruments, and outcomes, a narrative and thematic synthesis was adopted rather than meta-analysis, and direct evidence was distinguished from contextual evidence throughout. The terms systematic review, narrative synthesis, and thematic synthesis are used as defined in Section 4.9; PRISMA 2020 governs the reporting of the first, not the protocol in Section 3.2.

3.2 Protocol and Amendments

The protocol was not prospectively registered: PROSPERO accepts only health-related outcomes and was not applicable, the review was not registered in OSF, and no retroactive registration is claimed. A dated internal protocol was prepared at the outset (November 2026) covering the research questions, eligibility criteria, search plan, screening procedure, extraction variables, and appraisal instruments. Because it was internal, its amendments

were not externally audited before implementation; all four are disclosed here with timing, reason, and impact: forward citation chaining and a supplementary Crossref query (SS-15) were added after the string searches, adding 46 chained and 20 Crossref records to the screened pool with no criterion changed; the four-way evidence classification (direct, teaching-and-assessment, contextual, intervention) was formalised before full-text screening to keep school-level and adjacent-construct sources out of university claims, with no record added or excluded; a uniform rule for MMAT criterion 4.4 was adopted during appraisal because response-rate information was absent in most studies (Zenbil, 2021, criterion 4 set to cannot tell); and GRADE was adapted for narrative educational outcomes before certainty rating because pooled estimates and confidence intervals were unavailable (Section 4.8).

4. Methodology

4.1 Search Strategy: Sources, Strings, and Dates

The search distinguished five source types, which differ in indexing coverage: bibliographic databases (ERIC, Crossref, DOAJ, Dimensions); the Google Scholar search engine; ResearchGate, a sharing platform rather than a bibliographic database; Libyan university repositories and journal platforms (Zawia DSpace, and the platforms of the Al-Mukhtar Journal of Social Science, AlQalam-Hikma, the African Journal of Advanced Studies in Humanities and Social Sciences, Al-Qurtas, and the Tobruk, Sirte, and Misurata university journals); and full-text availability sites, used only to retrieve documents already identified. Scopus and Web of Science were not accessible (Section 8).

All platforms and databases were searched up to 14 November 2026. Fourteen English and Arabic strings derived from a core string were executed (SS-1 to SS-14, Appendix B), plus one supplementary Crossref query (SS-15). A counting cap applies: for each string, results across platforms were merged and de-duplicated, and the first twenty ranked results were screened; the search was therefore restricted rather than exhaustive, and records beyond the cap may have been missed (Section 8). Duplicates were removed by matching on author, year, and title, normalised for case and punctuation (58 in total). Records were also identified through citation chaining (Appendix B). Two verification searches on 14 November 2026 (a Crossref query restricted to 2024-2026, and an ERIC query combining Libya, listening, and higher education terms) identified no additional eligible record and changed no flow count in Table 5.

Table 2. Search sources, their type, dates, and screened records.

Source	Type	Date(s)	Strings	Records screened
Google Scholar	Academic search engine	1-14 Nov 2026	SS-1 to SS-14	20 per string (merged, de-duplicated)
ERIC	Bibliographic database	1-14 Nov 2026	SS-1 to SS-12	20 per string (merged)
Crossref	Bibliographic metadata database	14 Nov 2026	SS-15	20
DOAJ	Bibliographic database	Nov 2026	Core string	0 eligible records
Dimensions (free version)	Bibliographic database	Nov 2026	Core string	0 additional eligible records
ResearchGate	Academic sharing platform	1-14 Nov 2026	SS-1 to SS-12	20 per string (merged)
Zawia DSpace and Libyan journal platforms	Repositories and publishing platforms	1-14 Nov 2026	SS-1 to SS-14	20 per string (merged)
Verification searches (Crossref 2024-2026; ERIC Libya/listening/higher education)	Supplementary checks	14 Nov 2026	Not applicable	0 additional eligible records

Counting basis: the flow counts in Table 5 report records identified (14 strings x 20 merged-screened = 280, plus 20 from SS-15, plus 46 unique chained records = 346), not the total results retrieved, which was not systematically loggable because platform result lists overlap heavily and are not exportable in full.

4.2 Eligibility Criteria

Table 3. Inclusion and exclusion criteria, and handling rules for special source types.

Element	Inclusion criterion	Exclusion criterion
Population	English language learners or teachers in Libyan higher education	Pre-university populations only; non-Libyan populations
Phenomenon	Listening difficulty, listening comprehension processing, listening instruction or assessment	No listening component; general language difficulty without listening data
Context	Libyan universities and higher education institutions	School-only contexts (retained, if relevant, as contextual evidence only)
Study type	Empirical designs: descriptive, correlational, qualitative, quasi-experimental, mixed methods	News reports and non-scholarly outlets
Period	2015-2026	Outside the period
Language	Arabic or English	Other languages

Element	Inclusion criterion	Exclusion criterion
Publication	Verifiable academic sources (journals, repositories, theses)	Unverifiable or non-existent outlets
Outcomes	Listening difficulty measures, teaching/assessment practice data, or intervention effects	No reportable outcome of the above types

Handling rules: theses were eligible as empirical studies; articles in non-indexed national journals were eligible if the outlet and text were verifiable; conceptual articles and previous reviews were eligible only as contextual sources and were not appraised with MMAT; sources measuring adjacent affective constructs were eligible only as contextual evidence, and none measured listening anxiety with a listening-specific instrument; sources whose full text could not be obtained were appraised only on the available record, with every such judgement marked "cannot tell"; and sources reporting the same data were linked as multiple reports of one study (Section 4.5).

4.3 Screening and Selection

All records were screened by both researchers at title and abstract level against Table 3, then candidate reports were retrieved and assessed in full text, each receiving exactly one primary exclusion reason (full log in Appendix C). The original working screening log was not retained during preparation; the record was reconstructed by re-executing the fourteen documented search strings on the same platforms, filters, and date window, re-capturing the retrieved records under the same twenty-result cap, and re-screening after merging and de-duplication. Because platform ranking and coverage change over time, the reconstructed set cannot be guaranteed record-for-record identical to the original; the counts were anchored to the documented flow numbers rather than re-derived independently, and the reconstruction is reported as a verified approximation of the screening record, not as the record itself. Four safeguards were applied in place of independent review, none presented as a substitute for it: the criterion-fixed re-screening pass; external verification of every excluded entry as a real publication; page-anchored appraisal justifications (Appendix D); and a bidirectional citation audit with DOIs resolved against Crossref (Appendix A). These reduce, but do not remove, the risk of undetected error in one-person screening (Section 8).

During the revision stage, a criterion-based third assessment of the documented decisions was carried out by a an expert in statistical and methodology verification who had not participated in the original screening, extraction, or appraisal; and no linguistic or content judgement was delegated to him. Its scope was all 34 full-text eligibility decisions with their recorded reasons (19 excluded, 15 included); all 65 criterion-level MMAT judgements and all category assignments for the 13 empirical studies; and the extracted Omar and Pilus (2022) values, checked line by line against the retrieved article (seven checked, seven confirmed, no corrections). The eight outcome-level certainty ratings in Table 10 were re-checked for internal arithmetic consistency.

The unit of analysis differs by decision type, and the agreement statistics are reported separately for each type rather than pooled. For the eligibility audit, the unit is the full-text report decision, a binary Included/Excluded judgement applied once per report; raw agreement was 100% (34/34) and Cohen's kappa was 1.00, meaningful because both assessors used both categories and each report contributes one paired decision. Because the assessor saw the documented decisions and their justifications before concurring, the coefficient quantifies the criterion-consistency of the audit trail, not agreement between independent examiners. For the MMAT audit, no pooled kappa was computed across the 65 criterion judgements: the five criteria within each study are not statistically independent, sharing the same study and record, so pooling would overstate the effective sample size; agreement is reported descriptively instead (100% on each criterion, 13/13 paired judgements, and on all 13 category assignments). Because no disagreement arose, no consensus procedure was required, the final decisions are identical to the originals, and no PRISMA count changed.

Three qualifications govern these statistics. First, the assessment was a non-blinded audit of documented decisions; agreement statistics from such an audit describe the criterion-consistency of the audit trail, not inter-rater reliability under blinded conditions, and overstate the agreement that blinded dual screening would produce. Second, it was retrospective, conducted during revision — a verification procedure, not a feature of the original review design. Third, the paired decision sheets and the agreement computation were produced with documented tooling, and the audit cannot substitute for a third human reader screening the full texts blind.

4.4 Data Extraction

Data were extracted onto a structured form covering bibliographic details, setting and participants, design, instruments and reported psychometrics, variables measured, key quantitative results, intervention details, study-reported limitations, MMAT category and ratings, and the source locator for every value. One rule governed extraction: data not reported in the source were never recorded as facts, and no characteristic of a sample or instrument was inferred from a title or abstract; fields that could not be located in the consulted record were marked "not reported". The extraction matrix is in Appendix E. In the revision-stage third assessment (Section 4.3; Appendix G), the extracted Omar and Pilus (2022) values were checked against the retrieved article; seven checked, seven confirmed, no corrections.

4.5 Reports, Studies, and Independent Samples

Because two reports can describe one study, the review distinguishes three units of count: reports (published documents), studies (distinct investigations), and independent samples (non-overlapping participant groups). Overlap was checked for every pair of included sources, not assumed from titles alone. One overlap was identified and verified: Omar (2022) and Omar and Pilus (2022) both concern the same cohort of 62 learners at Sabratha University. No other overlap was identifiable from the available records; given single-site recruitment, undetected overlap cannot be excluded (Section 8.3).

Table 4. Linkage of reports to studies and independent samples, with the four-fold evidence classification.

Evidence category	Source (report)	Independent sample	How the evidence is used
1. Direct university-listening	Omar (2022)	Sabratha cohort (62)	Thematic findings
1. Direct university-listening	Othman (2025)	Ajdabiya (56)	Thematic findings
1. Direct university-listening	Mohammad & Alsharf (2025)	Azzaytuna (41)	Thematic findings
1. Direct university-listening	Bianku & Mahmud (2026)	Omar Al-Mukhtar (48)	Awareness finding, not proficiency
1. Direct university-listening	Zenbil (2021)	Zawia students (40)	Thematic findings (test-based strand)
1. Direct university-listening	Fatah & Farj (2024)	Almarj first-semester cohort (30)	Thematic findings (lexical self-report, exposure)
2. Direct teaching-assessment	Al-shareef (2022)	Zawia teachers (6 interviewed)	Teaching and assessment findings; not a substitute for student outcomes
3. Contextual	Pathan et al. (2016)	35 teachers, 12 schools	Background on school preparation
3. Contextual	Alghanimi & Ghazalla (2021)	School teachers and classrooms	Washback mechanism; school level
3. Contextual	Aldarasi (2020)	22 students, 8 teachers	Speaking-centred anxiety; indirect for listening
3. Contextual	Halali et al. (2026)	Misurata (97)	Adjacent construct; no listening-anxiety prevalence
3. Contextual	Bin Naji et al. (2026)	Not applicable (50 articles)	Comparison and interpretation only
3. Contextual	Bahron & Abujela (2026)	Not applicable	Argument and interpretation only
4. Intervention	Omar & Pilus (2022)	Sabratha cohort (62; same sample as Omar, 2022)	Tested in Section 5.7; never counted as confirmation of baseline
4. Intervention	Aldizeeri (2025)	Al-Marqab (10)	Tested in Section 5.7

The corpus therefore contains 15 reports, 14 studies, and 12 independent samples: six higher education samples bearing directly on listening difficulty among university students, one teaching-assessment sample, two intervention samples (one being the Sabratha cohort already counted), and four further samples supplying contextual evidence only.

4.6 Quality Appraisal (MMAT Version 2018)

Methodological quality was appraised with the original Mixed Methods Appraisal Tool, version 2018 (Hong et al., 2019): two screening questions and five design categories, each with five criteria. No separate correlational category exists in the tool; correlational studies were classified according to their actual design, with the descriptive category applied to cross-sectional surveys without intervention and the justification recorded per study. Responses were Yes, No, or Can't tell, each justified against the consulted record; no aggregate score was computed, in line with the tool's guidance. Two sources (one narrative review, one conceptual article) are not appraisable with MMAT and are assessed narratively. The full criterion-level records are in Appendix D; Table 7 summarises them. A criterion-based third assessment of all 65 criterion judgements and all category assignments during the revision stage (Section 4.3; Appendix G) concurred with every judgement; agreement is reported descriptively rather than as a pooled kappa, and no rating changed.

4.7 Risk of Bias

Risk of bias was assessed per study against domains appropriate to its design, distinguishing three related but not interchangeable concepts: methodological quality (Section 4.6) is criterion-level appropriateness of design and execution; risk of bias is the likelihood that conduct or analysis systematically distorted a study's results; and certainty of evidence (Section 4.8) is confidence in a synthesised outcome, which can be lower than any single study's quality. For descriptive and correlational surveys the domains were selection, measurement (including self-report common-method variance), and reporting; the intervention studies added confounding and absence of blinding; the qualitative strand of the mixed-methods studies added participant selection, data collection adequacy, and researcher influence. Publication bias was assessed at the level of the body of evidence (Section 5.4). Table 8 reports the study-level judgements with the basis of each.

4.8 Certainty of Evidence (GRADE)

Certainty was rated with GRADE (Guyatt et al., 2008) at the level of precise outcomes rather than broad themes, because the studies measure different variables and a wide theme would mix incommensurable results. Eight

outcomes were rated: perceptual and phonological difficulties during listening; lexical gaps during listening; listening task anxiety; metacognitive awareness of listening; teaching and assessment neglect of listening; severity of material and environmental constraints; effectiveness of connected speech training; and effectiveness of cooperative video-based learning.

The suitability of GRADE for this corpus was assessed before application rather than assumed. It was retained over an unstructured narrative rating because it forces every judgement standing behind a conclusion to be stated, located, and auditable; it was not applied as an official GRADE exercise because its quantitative conventions (pooled estimates, confidence intervals, effect sizes) do not exist for these outcomes. The application is an explicitly declared adaptation for descriptive educational outcomes: no trial-designed risk-of-bias instrument was used, the observational starting point was applied to every outcome, and the ratings are one rater's documented judgements (Section 8). Because no study meets the standard of an adequately described randomised trial (Section 5.4), every outcome starts at low certainty, a conservative choice applied uniformly.

Each of the five domains (risk of bias, inconsistency, indirectness, imprecision, publication bias) was judged explicitly: a domain judged serious lowers the rating by one level, and every downgrade, and every decision not to downgrade, is justified in Table 10, so the final level follows arithmetically from the domain judgements. Inconsistency judgements distinguished differences in instruments, outcome definition, direction, magnitude, and non-commensurability; where constructs cannot be compared at all, the problem is recorded as one of comparability rather than of inconsistent direction. Independent samples, not reports, were counted; contextual sources were never entered as confirmatory evidence. A fixed imprecision rule governed the small-sample outcomes: in the absence of pooled estimates and confidence intervals, imprecision was judged narratively on the number of independent samples, their sizes, and the convergence of their direction (Table 10). The resulting levels express confidence in the direction and consistency of each finding, not statistical precision; no magnitude is implied. The eight ratings were re-checked for internal arithmetic consistency in the second assessment (Section 4.3; Appendix G).

4.9 Synthesis Method

Findings were synthesised thematically following the three-stage logic of thematic synthesis (line-by-line coding, grouping into descriptive themes, developing analytical themes) described by Thomas and Harden (2008). Coding was performed on the findings and discussion sections of the included sources; direct perceptual evidence was synthesised separately from contextual and intervention evidence; and the themes were organised using the four-factor taxonomy (speaker, listener, message, environment) supplemented by pedagogical and systemic categories that emerged inductively.

5. Results

5.1 Study Selection and PRISMA Flow

Table 5. PRISMA 2020 flow of records, reports, studies, and independent samples.

Stage	n
Records identified from string searching (14 strings x 20 merged-screened)	280
Records identified from supplementary Crossref query (SS-15)	20
Records identified from citation chaining (unique)	46
Records identified in total	346
Duplicates removed before screening	58
Records screened at title/abstract level	288
Records excluded at title/abstract level	254
Reports sought for retrieval	34
Reports not retrievable	0
Reports assessed in full text	34
Full-text reports excluded, with reasons	19
Reports included	15
Studies (after linking the two Sabratha reports)	14
Independent samples	12

The 346 figure counts records identified under the restricted search and capping rules, not the total retrieved (Section 4.1). Of the 19 full-text exclusions, 6 reports addressed other skills only, 5 other countries, 3 pre-university contexts without direct relevance to university preparation, 2 non-scholarly outlets, 2 populations other than Libyan university learners, and 1 was a duplicate report of data already represented. One report (Omar & Pilus, 2022) was initially appraised from the publisher's abstract; its full text was retrieved in the final verification pass and its appraisal completed from the full text (Sections 4.6, 5.7). The record-by-record log is in Appendix C, where the reconstruction of the original log is disclosed.

5.2 Characteristics of the Included Studies

Table 6. Characteristics of the included reports (n = 15). Zenbil (2021) is mixed-methods by design and is appraised in the quantitative descriptive MMAT category because its two strands are analysed separately (Section 5.4).

No.	Study	Setting	Sample	Design	Instrument(s)	Psychometrics	Key findings
1	Pathan et al. (2016)	Libyan schools	35 teachers, 12 schools	Survey	Teacher questionnaire	Not reported	Training, coordination, infrastructure deficits
2	Aldarasi (2020)	Libyan university students	22 students, 8 teachers	Qualitative	Semi-structured interviews	n/a	Anxiety tied to accents, settings, comparison
3	Alghanimi & Ghazalla (2021)	Public schools, Tripoli	Teachers and classrooms	Quantitative	Questionnaire, observation	Not reported	Listening materials ignored; examination washback
4	Zenbil (2021)	University of Zawia	40 first-year students, 30 teachers	Mixed methods	Listening test, questionnaire	Not reported	Retrieval, recognition, substitution; materials gaps
5	Omar (2022)	Sabratha University	62 third-year majors, 90% female	Quantitative descriptive	LCPQ (38 items)	alpha = 0.924	Perception 3.14, parsing 3.25, utilisation 3.21; phase r = .789-.842
6	Omar & Pilus (2022)	Sabratha University (same cohort)	62 learners aged 19-21; treatment 32, control 30, four intact classes	Quasi-experimental (pretest-posttest)	Adapted LCPQ, pre/post	alpha = 0.90/0.89/0.89	Large effects after 12 weeks, 24 sessions (d = 1.14-1.19)
7	Al-shareef (2022)	University of Zawia	Teachers of listening courses (6 interviewed)	Mixed methods	Questionnaire, interviews	Not reported	Objectives rarely clarified; scoring, tasks, texts weak
8	Fatah & Farj (2024)	Benghazi (Almarj branch)	30 first-semester majors	Mixed methods	Questionnaire, interviews, tasks	Not reported	76% limited vocabulary; 70% no practice; 50% daily input
9	Othman (2025)	Ajdabiya University	56 undergraduates, 89.3% female	Quantitative descriptive	Five-domain questionnaire	Partial	55.4% never took the listening course; 48.2% fear not understanding
10	Mohammad & Alsharf (2025)	Azzaytuna University	41 undergraduates, semesters 4-8	Quantitative descriptive	22-item Likert questionnaire	alpha = 0.881	Physical setting most severe (M = 4.30); vocabulary highest (M = 4.39); overall M = 3.76
11	Aldizeeri (2025)	Al-Marqab University	10 students	Quasi-experimental	Pre-post listening scores, questionnaire	Not reported	Cooperative YouTube improved comprehension and motivation
12	Halali et al. (2026)	Misurata University	97 undergraduates	Quantitative descriptive	Online questionnaire	Partial	Vocabulary anxiety moderate (M = 2.7951), during listening and speaking
13	Bianku & Mahmud (2026)	Omar Al-Mukhtar University	48 third/fourth-year majors	Quantitative descriptive	MALQ (21 items)	Validated instrument	Problem-Solving highest (5.11/6); Directed Attention 4.45, Planning-Evaluation 4.46; Mental Translation 5.06
14	Bin Naji et al. (2026)	Libyan higher education	50 articles (2022-2026)	Narrative review	Reflexive thematic analysis	n/a	Teacher-centred pedagogy, exam-oriented curricula, anxiety, memory overload, weak self-regulation
15	Bahron & Abujela (2026)	University of Zawia	Not applicable	Conceptual	Argumentative analysis	n/a	Listening overlooked and tested rather than taught

Table 6 reports the condensed characteristics; Appendix E records the full matrix with page-level locators for every value. No field was filled by inference; unreported values are marked as such.

5.3 Methodological Characteristics of the Evidence

Four features characterise the evidence base. The empirical studies are geographically dispersed across western, central, and eastern Libya (Zawia, Sabratha, Al-Marqab, Azzaytuna, Misurata, Benghazi-Almarj, Ajdabiya, Omar Al-Mukhtar), but the southern universities and private institutions are effectively absent. The dominant design is the cross-sectional descriptive questionnaire (7 of 13 empirical studies); only two studies test an intervention, and none is longitudinal. Sample sizes range from 10 to 97, and most samples are heavily female-skewed (90% in Omar, 2022; 89.3% in Othman, 2025). Psychometric documentation is uneven: three studies report reliability coefficients, and none a full instrument validation.

5.4 Methodological Quality and Risk of Bias

Table 7. MMAT (2018) summary by study (full criterion-level records in Appendix D). SQ1/SQ2 = screening questions; Y = yes, N = no, CT = cannot tell. Categories follow each study's allocation or design mechanism, not the authors' self-label; Zenbil (2021) is appraised in the quantitative descriptive category because its two strands are analysed separately (Section 5.4). A retrospective second assessment concurred with all 65 criterion judgements and all category assignments (Section 4.3; Appendix G).

Study	MMAT category	SQ1	SQ2	C1	C2	C3	C4	C5
Pathan et al. (2016)	Quantitative descriptive	Y	Y	Y	CT	Y	CT	Y
Aldarasi (2020)	Qualitative	Y	Y	Y	Y	Y	Y	CT
Alghanimi & Ghazalla (2021)	Quantitative descriptive	Y	Y	Y	CT	Y	CT	Y
Zenbil (2021)	Quantitative descriptive	Y	Y	Y	CT	Y	CT	Y
Omar (2022)	Quantitative descriptive	Y	Y	Y	CT	Y	CT	Y
Omar & Pilus (2022)	Randomised (cluster allocation)	Y	Y	CT	Y	N	CT	CT
Al-shareef (2022)	Mixed methods	Y	Y	Y	Y	Y	N	Y
Fatah & Farj (2024)	Mixed methods	Y	Y	Y	CT	Y	N	CT
Othman (2025)	Quantitative descriptive	Y	Y	Y	CT	CT	CT	Y
Mohammad & Alsharf (2025)	Quantitative descriptive	Y	Y	Y	CT	Y	CT	Y
Aldizeeri (2025)	Non-randomised	Y	Y	N	Y	Y	N	Y
Halali et al. (2026)	Quantitative descriptive	Y	Y	CT	CT	CT	CT	Y
Bianku & Mahmud (2026)	Quantitative descriptive	Y	Y	Y	N	Y	CT	Y
Bahron & Abujela (2026)	Not appraisable (conceptual)	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Bin Naji et al. (2026)	Not appraisable (narrative review)	n/a	n/a	n/a	n/a	n/a	n/a	n/a

The category assignments follow each study's actual design as reported in its methods section, not its title or abstract label. Omar and Pilus (2022) reports that its four intact classes were randomly assigned as whole groups to treatment (n = 32) and control (n = 30) at cluster level (pp. 71-72), while subject selection was not randomised (p. 77) and the design is self-identified as quasi-experimental; because allocation to conditions was randomised at cluster level, the study is appraised in the randomised category, with the random-assignment criterion judged cannot tell (the procedure is not described, only four clusters were allocated; Appendix D). Selection and allocation are reported as separate properties throughout. The two cross-sectional analytical surveys (Omar, 2022; Bianku & Mahmud, 2026) are placed in the quantitative descriptive category because they are surveys without intervention; the tool contains no separate correlational category. The two mixed methods studies differ in one criterion: Al-shareef (2022) reports the two strands but not their integration; Fatah & Farj (2024) neither integrates the strands nor addresses divergences between them.

Table 8. Risk of bias by study, tied to design, with the basis of each judgement.

Study (design)	Selection	Measurement	Reporting	Design-specific concerns	Basis
Pathan et al. (2016; survey)	High (schools not randomly selected)	Moderate (unvalidated questionnaire)	Moderate (partial response-rate reporting)	n/a	Article methods
Aldarasi (2020; qualitative)	Moderate (purposive, justified)	Low (interview triangulation)	Moderate (reflexivity not addressed)	Researcher influence not documented	Thesis methods
Alghanimi & Ghazalla (2021; survey + observation)	High (single city)	Moderate	Moderate	Observation protocol not published	Article methods
Zenbil (2021; test + questionnaire)	High (single site, intact groups)	Moderate (test psychometrics unreported)	Moderate	Common-method and test-form effects unquantified	Thesis methods
Omar (2022; descriptive survey)	High (single site, 90% female)	Low (validated 38-item instrument)	Low (transparent statistics)	Common-method variance	Article methods
Omar & Pilus (2022; intervention)	High (non-random selection, p. 77; cluster allocation of four classes, undescribed procedure, pp. 71-72)	Low-moderate (validated adapted LCPQ, alpha 0.89-0.90; self-report)	Moderate (pre-test missing cases eliminated; final n unstated)	No blinding or fidelity check	Full text (pp. 71-77, 85)
Al-shareef (2022; mixed methods)	High (single site)	Moderate	Moderate (questionnaire psychometrics unreported)	Teacher-side data only	Thesis methods
Fatah & Farj (2024; mixed methods)	High (single site, 30 students)	Moderate	High (no psychometric or audit-trail reporting)	Small sample; integration under-documented	Full text
Othman (2025; descriptive survey)	High (single site)	Moderate (psychometrics partial)	Low	Common-method variance	Article methods

Study (design)	Selection	Measurement	Reporting	Design-specific concerns	Basis
Mohammad & Alsharf (2025; descriptive survey)	High (single site)	Low (alpha = 0.881 reported)	Low	Common-method variance	Article methods
Aldizeeri (2025; quasi-experiment)	High (n = 10, no randomisation)	Moderate (unvalidated pre-post scores)	Moderate	Confounding uncontrolled at n = 10	Article methods
Halali et al. (2026; descriptive survey)	Unclear (sampling not fully described)	Moderate (psychometrics partial; adjacent construct)	Moderate	Construct indirectness	Abstract/ journal record
Bianku & Mahmud (2026; cross-sectional)	High (purposive single site)	Low (validated MALQ)	Moderate	Correlational design; no causal inference	Study report

Reconciling study-level risk of bias with outcome-level certainty: a serious study-level limitation does not mechanically downgrade every outcome that cites the study; the GRADE domain is judged for the body of evidence behind each outcome, considering how much the study contributes, whether the outcome rests on it alone, and whether the problem threatens the direction of the finding or only its precision. For perceptual and phonological difficulties, four independent samples contribute, the instruments are validated or transparently reported, and the weaknesses bear on transferability and precision rather than direction, so risk of bias was not downgraded; the same weaknesses remain registered in Table 8 and in the imprecision judgements. For metacognitive awareness, a single study carries the entire outcome, so its limitations transfer directly and the relevant domains were downgraded. For teaching and assessment neglect, the teacher-side self-report with unreported psychometrics threatens direction as well as precision, so risk of bias was downgraded.

At body-of-evidence level, publication bias could not be quantified: the eligible national literature is small, concentrated in a limited set of national outlets, and dominated by English-language publications, so absence of evidence for publication bias is not evidence of its absence. Selective outcome reporting within studies could not be audited because most did not publish instrument protocols or analysis code.

5.5 Thematic Synthesis of Direct Evidence

The direct evidence supports six themes. Every figure is tied to its study, sample, instrument, and locator; none is generalised beyond its sample, and perceived difficulty and measured performance are never merged.

Theme 1: Perceptual and phonological difficulties with connected speech. The most consistently documented cluster (Table 9): at Sabratha University, the 38-item LCPQ administered to 62 third-year majors produced an overall mean of 3.20 (SD = 0.75), with speed of delivery the highest-rated item after anxiety (M = 3.56, SD = 1.03) and the connected-speech features of assimilation, linking, omission, and reduction at M = 3.45 (SD = 1.25) (Omar, 2022, pp. 13-14); these are questionnaire results. The test-based strand converges: at the University of Zawia, the listening test showed that students "lack some cognitive skills and strategies such as retrieving and recognition and substitution strategy" (Zenbil, 2021, Abstract section); at Ajdabiya University, 37.5% of 56 students reported difficulty recognising a known word because of the speaker's pronunciation (Othman, 2025, p. 183); and at Azzaytuna University, unfamiliar accents and high-speed speech ranked among the material-related difficulties (Mohammad & Alsharf, 2025, p. 2148). The questionnaire and test strands measure perceived difficulty and observed performance respectively, reported side by side, not merged.

Theme 2: Lexical gaps during listening. In the Benghazi-Almarj sample of 30 first-semester majors, "76% of the participants agreed they have a limited vocabulary, while 70% of the participants do not generally practice listening" (Fatah & Farj, 2024, Abstract section). At Azzaytuna, limited vocabulary was the highest single item in that instrument (M = 4.39, SD = 0.92; Mohammad & Alsharf, 2025, p. 2150); at Sabratha, the new-words-and-expressions item scored 3.55 and texts containing idioms and jargon attracted the highest utilisation mean of 3.60 (Omar, 2022, pp. 14, 17); at Ajdabiya, 48.2% of students reported difficulty with texts containing many unfamiliar words (Othman, 2025, p. 182). These four independent samples used four different instruments, and the 76% figure is a property of one sample of 30 first-semester students, not a national prevalence rate.

Theme 3: Cognitive processing, memory, and metacognitive limitations. The Sabratha parsing evidence shows the highest means for connected speech on an unfamiliar topic (M = 3.79), long and complex sentences (M = 3.69), fatigue during long texts (M = 3.65), and large amounts of new information in a short time (M = 3.63) (Omar, 2022, p. 16). At Ajdabiya, 46.4% of students reported stopping to think about an unknown word and losing the continuation of the input, and 42.9% found real-time inference of unknown words difficult (Othman, 2025, p. 183). The metacognitive profile at Omar Al-Mukhtar University (validated MALQ, 48 majors) shows Problem-Solving highest (M = 5.11 of 6), Directed Attention lowest (M = 4.45), Planning-Evaluation low (M = 4.46), and heavy reliance on Mental Translation (M = 5.06) (Bianku & Mahmud, 2026, Abstract section). Two cautions apply: MALQ measures awareness of strategies, not listening proficiency; and the phase correlations (r = .789-.842; Omar, 2022, p. 19) are evidence of association, not of a causal path.

Theme 4: Affective barriers. The single highest-rated item in the Sabratha instrument was emotional: "I feel nervous and worried when I do not understand the connected speech" (M = 3.94, SD = 1.28) (Omar, 2022, p. 14), and 48.2% of the Ajdabiya sample reported fearing before a listening task that they would not understand (Othman, 2025, p. 183). No study measured listening anxiety with a dedicated listening-anxiety scale; the adjacent evidence (vocabulary anxiety at Misurata, M = 2.7951, Halali et al., 2026; speaking-centred anxiety in Aldarasi, 2020, both Abstract sections) is reported in Section 5.6 and is not used to estimate the prevalence of listening anxiety. Whether anxiety depresses performance, poor performance generates anxiety, or both, has not been tested in this setting.

Theme 5: Pedagogical and assessment neglect. The teaching and assessment evidence is teacher-side and curricular, reported separately from student outcomes. At the University of Zawia, teachers "always set goals for listening tasks" while "most of them rarely clarify the objectives of the anticipated listening tasks to the students", and reported weaknesses in scoring, task setting, text selection, experience, resources, and training (Al-shareef, 2022, Abstract section); teachers at the same university reported "lack of authentic materials, lack of equipments or teaching aids, dependence on ready-made materials" and a need for training (Zenbil, 2021, Abstract section). On the curricular side, 55.4% of the Ajdabiya sample had never studied the Listening Concepts course (Othman, 2025, pp. 181-182). That exclusion of listening from examinations produces neglect of its teaching is supported directly at school level, where exclusion was followed by "a wilful neglect of teaching this skill" (Alghanimi & Ghazalla, 2021, Abstract section); at university level the same mechanism is a plausible interpretation, not a demonstrated causal relationship.

Theme 6: Material, technological, environmental, and systemic constraints. In the Azzaytuna study, physical-setting difficulties formed the most severe domain of that instrument (M = 4.30, SD = 0.91), higher than listener-related (M = 3.60) and material-related (M = 3.38) difficulties, within an overall mean of M = 3.76 (SD = 1.07); the domain covers poor audio quality, environmental noise, and crowded classrooms (Mohammad & Alsharf, 2025, pp. 2148, 2150). This finding is a property of one sample of 41 students; no national ranking of difficulty sources follows from it. The exposure profile is documented at Benghazi-Almarj, where only half of the 30 participants reported daily English input (Fatah & Farj, 2024, Abstract section), and the Ajdabiya study reports that listening problems affected both language learning and overall academic performance (Othman, 2025, p. 168).

Table 9. Thematic synthesis of the direct evidence (independent samples; the Sabratha cohort counted once). Perceived difficulty and measured performance are reported separately and never merged.

Theme	Direct supporting samples	Key results	Evidence limits	Certainty (Table 10)
1. Perceptual and phonological difficulties	Omar (2022); Zenbil (2021, test strand); Othman (2025); Mohammad & Alsharf (2025)	Speed M = 3.56; connected-speech M = 3.45; 37.5% misrecognise known words; test-based deficits	Single-institution samples; questionnaire and test results separate; no standardised measure	Low
2. Lexical gaps	Fatah & Farj (2024); Omar (2022); Othman (2025); Mohammad & Alsharf (2025)	76% (n = 30, one site); M = 4.39 highest item; idiom-text utilisation M = 3.60	Percentage and Likert metrics not commensurable; self-report	Very low
3. Cognitive processing, memory, metacognition	Omar (2022); Zenbil (2021); Othman (2025); Bianku & Mahmud (2026)	Parsing means 3.63-3.79; r = .789-.842 among phases; MALQ: Directed Attention 4.45, Mental Translation 5.06	Self-report proxies; MALQ measures awareness, not proficiency; correlations not causal	Very low
4. Affective barriers	Omar (2022); Othman (2025)	Anxiety item M = 3.94 (highest single item in that instrument); 48.2% fear not understanding	No dedicated listening-anxiety instrument in any study	Very low
5. Pedagogical and assessment neglect	Al-shareef (2022); Zenbil (2021); Othman (2025); Alghanimi & Ghazalla (2021, school level)	Objectives set but rarely clarified; 55.4% never took the listening course; school-level washback	Teacher-side and school-level data; university washback interpretive	Very low
6. Material, technological, environmental constraints	Mohammad & Alsharf (2025)	Physical setting most severe domain (M = 4.30)	One quantitative source (n = 41); no national ranking follows	Very low

5.6 Contextual Evidence

Six sources provide contextual support and are not counted as evidence for any prevalence claim. The school-level studies explain the preparation students bring to university: teachers reported insufficient training, weak coordination, and infrastructure shortages (Pathan et al., 2016, p. 19), and classroom observation in Tripoli found that listening materials in the course books were "ignored and skipped" under examination washback (Alghanimi & Ghazalla, 2021, Abstract section). The anxiety-adjacent studies clarify what has and has not been measured: Aldarasi (2020) documents foreign language anxiety bound up with accents, classroom settings, and social comparison, a speaking-centred construct that only partially overlaps listening; the Misurata study measured vocabulary learning anxiety (M = 2.7951, "particularly during listening and speaking") with a vocabulary-anxiety

instrument, not a listening-anxiety scale (Halali et al., 2026, Abstract section). The narrative review of Bin Naji et al. (2026) synthesises 50 recent articles, identifying teacher-centred pedagogy, exam-oriented curricula, anxiety, working-memory overload, and weak self-regulation; it is used for comparison, not as primary evidence. The conceptual article of Bahron and Abujela (2026) argues that listening has been "overlooked and seen as a passive skill" and tested rather than taught (Abstract section); it contributes argument, not data.

5.7 Intervention Evidence

Table 13 (Appendix H) records both studies' designs, samples, results, effect sizes, and limits of evidence.
Neither study warrants a claim of proven effectiveness. The connected-speech result concerns reported problems, not measured proficiency; the cooperative-learning result is performance-based but rests on ten participants without randomisation. Both indicate that listening outcomes in this setting respond to instructional design, and both require replication with randomised assignment, standardised proficiency outcomes, adequate samples, and reported confidence intervals before any programme is scaled.

5.8 Certainty of the Evidence

Table 10. GRADE assessment at outcome level. Starting level: low for all outcomes (observational starting point applied conservatively; Section 4.8). NS = not serious; S = serious (one-level downgrade); U = undetected but not assessable (no downgrade; flagged). The final level follows arithmetically: low minus one downgrade = very low. Levels express confidence in the direction and consistency of narrative findings, not statistical precision; no magnitude is implied.

Table 14 (Appendix I) summarises the six contextual sources, their samples, and how each is used.

Outcome	Direct sources (independent samples)	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias	Final certainty
Perceptual and phonological difficulties during listening	Four samples (Table 9, Theme 1)	NS: validated or transparent instruments	NS: convergent direction across four samples	NS: all four measure perceptual/phonological difficulty; connected speech from one sample	NS: four samples (n = 30-62) converge in direction	U	Low
Lexical gaps during listening	Four samples (Table 9, Theme 2)	NS: instruments transparently reported	NS: direction converges; differing metrics are a limitation, not inconsistent direction	NS	S: four small samples (n = 30-62), self-report only, no CIs	U	Very low
Listening task anxiety	Two samples (Table 9, Theme 4)	NS: items transparently reported	NS: consistent direction	S: no dedicated listening-anxiety instrument in any study	NS	U	Very low
Metacognitive awareness of listening	One study (Table 9, Theme 3)	NS: validated MALQ	NS: single study	S: MALQ measures awareness, not proficiency	S: single study, one site, n = 48	U	Very low
Teaching and assessment neglect of listening	Three samples (Table 9, Theme 5)	S: teacher-side self-report; psychometrics unreported	NS	S: measures practice, not listening outcomes	NS	U	Very low
Effectiveness of connected speech training	Omar & Pilus (2022)	S: cluster allocation of four classes, undescribed procedure, non-random selection (pp. 71-72, 77); no blinding; incomplete outcome data	NS: single trial	S: perception outcomes, not standardised proficiency	S: single trial, reduced analysed n, no CIs	U	Very low
Severity of material and environmental constraints	Mohammad & Alsharf (2025)	NS: reliability reported (alpha = 0.881); single-study limits carried by imprecision	NS: single study	NS: physical-setting difficulty measured directly for listening	S: single study, one site, n = 41, no CIs	U	Very low
Effectiveness of cooperative video-based learning	Aldizeeri (2025)	S: no randomisation; confounding uncontrolled	NS: single study	NS: performance-based outcome	S: n = 10, no CIs	U	Very low

Imprecision judgements record the limits of the samples behind each outcome: single studies, small institutionally concentrated samples, and absent confidence intervals; no pooled estimate was attempted. Publication bias is recorded as U (undetected but not assessable) for every outcome rather than downgraded: the corpus sits in

national and regional outlets that indexing-based detection does not cover, no outcome rests on enough studies for asymmetry to be testable, and downgrading on unassessable grounds would be arbitrary; U flags that gap explicitly. Indirectness downgrades were applied where the construct measured differs from the outcome rated (awareness rather than proficiency; practice rather than listening outcomes; perception rather than standardised performance). Contextual sources (Section 5.6) were never entered as confirmatory evidence or used to raise certainty. No domain was upgraded: the corpus offers no large, consistent, replicated effects.

5.9 An Integrated Framework of Listening Challenges in Libyan Higher Education

The framework organises the evidence into five levels — inputs (acoustic and instructional), processing (perception, parsing, utilisation, constrained by working memory and attention), regulators (metacognitive strategies, anxiety, confidence, language experience), outputs (comprehension, word recognition, sustained follow-up of meaning, academic performance), and an institutional level (curricula, teaching practices, assessment policy, teacher education, infrastructure) acting on every other level — and states, for each link, whether the corpus supports it or the link remains a hypothesis.

Evidence-supported links: input characteristics to perception failure (speed M = 3.56, connected-speech features M = 3.45; Omar, 2022, p. 14); perception-parsing-utilisation propagation (r = .789-.842; Omar, 2022, p. 19); lexical gaps to recognition and utilisation failure (four independent samples; Section 5.5, Theme 2); assessment policy to neglect of teaching (school-level washback; Alghanimi & Ghazalla, 2021); infrastructure to difficulty (physical setting M = 4.30; Mohammad & Alsharf, 2025). Hypothesis links: anxiety to reduced performance (co-occurrence only); metacognitive awareness to better performance (MALQ measures awareness, not proficiency); strategy instruction to improved outcomes (untested in Libya). The framework is an organising interpretation, not a causal model verified in Libyan universities.

6. Discussion

6.1 Principal Findings

Four challenges recur across the direct university evidence: perceptual difficulty with connected speech, lexical gaps during listening, interdependent cognitive-processing demands, and affective barriers; pedagogical, assessment, and environmental constraints frame all four. Each is stated at the certainty the evidence supports (Table 10) — low for the perceptual finding, very low for the rest — and none is generalised beyond its sample.

6.2 Perceptual and Phonological Challenges

The recurrence of perceptual difficulty across four samples is directly supported; its explanation is not. Three plausible accounts, documented separately and not tested jointly, are theory-consistent rather than demonstrated: print-based vocabulary learning, where the phonological modifications of connected speech distort the citation forms learners know from print (Mohammad & Alsharf, 2025, p. 2148; Omar, 2022, p. 20); limited exposure, with only half of the Benghazi-Almarj sample reporting daily English input (Fatah & Farj, 2024, Abstract section); and instructional neglect, with school listening materials skipped under examination washback (Alghanimi & Ghazalla, 2021, Abstract section). The intervention evidence is consistent: when decoding was trained explicitly for twelve weeks, reported problems across all three phases fell with large effect sizes (Omar & Pilus, 2022). These are convergent interpretations, not tested mechanisms.

6.3 Lexical and Linguistic Constraints

The lexical link is the most directly documented in the corpus: the highest-rated perception item after speed was the density of new words in connected speech (M = 3.55; Omar, 2022, p. 14), and the highest-rated utilisation item was texts containing idioms and jargon (M = 3.60; p. 17). It is directly supported at the level of reported difficulty and remains an interpretation at the level of cause: no study measured vocabulary size with a test and related it to measured listening performance, so the direction of the relationship cannot be established from this corpus.

6.4 Cognitive and Metacognitive Processing

The parsing evidence and the MALQ profile fit one account: learners compensate through inference and translation precisely where bottom-up decoding and sustained attention fail. The phase correlations (r = .789-.842; Omar, 2022, p. 19) support treating the phases as interdependent, but they are associations, not a demonstrated causal path, and MALQ scores measure strategic awareness rather than proficiency; any claim that strategy training would improve performance is a hypothesis for intervention research, not a finding of this review.

6.5 Affective Factors and Listening Anxiety

The affective evidence shows co-occurrence, not causation: the highest-rated single item in the Sabratha instrument is an anxiety item (M = 3.94; Omar, 2022, p. 14), and 48.2% of the Ajdabiya sample feared not understanding (Othman, 2025, p. 183), yet no study used a dedicated listening-anxiety instrument, so the prevalence of listening anxiety in Libyan universities is unknown. The cooperative-learning study reports increased confidence under its condition (Aldizeeri, 2025, p. 1433), showing that affect is responsive to pedagogy within that study; the causal direction remains untested and could be reciprocal or confounded by proficiency.

6.6 Pedagogical and Assessment Practices

The teaching and assessment findings explain why the learner-internal difficulties persist. Teachers set goals but rarely clarify objectives and report deficits in scoring, task setting, text selection, resources, and training (Al-shareef, 2022; Zenbil, 2021, both Abstract sections); more than half of one sample had never taken the dedicated listening course (Othman, 2025, pp. 181-182). The school-level washback mechanism (Alghanimi & Ghazalla, 2021) and the university-level assessment weaknesses (Al-shareef, 2022) together identify examination policy as a promising lever for reform, one worth piloted study and trial; at university level this remains an interpretation of practice data, not a tested causal claim.

6.7 Environmental and Institutional Constraints

The severity of physical-setting difficulty in the Azzaytuna sample (M = 4.30, the most severe domain of that instrument; Mohammad & Alsharf, 2025, p. 2150) locates part of the documented difficulty in infrastructure rather than learner language. The finding comes from one sample of 41 students and is not a national ranking; it identifies a remediable factor that university administrations control directly, and it challenges the assumption that listening interventions should target learners first.

6.8 Evidence from Educational Interventions

Neither intervention study, alone or together, establishes effectiveness: one randomly allocated four intact classes at cluster level with an undescribed procedure while selecting participants non-randomly and measuring perceptions (Omar & Pilus, 2022); the other had ten participants and no randomisation (Aldizeeri, 2025). Their value is directional — outcomes in this setting respond to instructional design — and their combined weakness defines the design standards future Libyan intervention research must meet.

6.9 Comparison with International Evidence

The Libyan profile matches the canonical international patterns: the three-phase distribution of processing problems parallels the pattern reported for other EFL populations (Goh, 2000), and the four-factor cause structure corresponds to the taxonomy reported for Saudi EFL learners, whose difficulties included "pronunciation, speech rate, vocabulary, speaker accent, inability to concentrate, stress, and forgetfulness" (Hamouda, 2013, as cited in Othman, 2025, p. 179). Two features nonetheless distinguish the Libyan evidence: examination-washback suppression of an entire skill at the level of national school-examination policy, and the prominence of infrastructure-level difficulty in a low-resource setting. International findings are used for comparison, not as proof that the same magnitudes hold in Libya.

6.10 Theoretical and Practical Implications

Theoretical. The corpus supports treating the three-phase model as recursive and interdependent rather than sequential, and indicates that the four-factor taxonomy requires a fifth, institutional layer that the taxonomy does not name; this layer is an organisational finding extracted from the evidence, not an experimentally demonstrated causal discovery. **Pedagogical and curricular.** Listening should be taught as a trainable process, with explicit decoding and strategy instruction; the Listening Concepts course and connected-speech content should be guaranteed and resourced. **Assessment.** Placing listening in an appropriate position within assessment may be worth study and trial; the causal effect has not been tested, and any reform should be piloted with clearly specified listening tasks before system-wide adoption. **Institutional.** Audio infrastructure and teacher education are preconditions for the pedagogical reforms. Each implication is bounded by the certainty ratings in Table 10.

6.11 Research Gaps and Future Directions

The synthesis identifies ten gaps, converted into a research agenda in Table 12. The most consequential are the absence of standardised proficiency measurement alongside perceptions, the near-absence of experimental designs, the absence of dedicated listening-anxiety measurement, and the lack of longitudinal tracking of listening development across degree programmes.

7. Implications and Recommendations

7.1 Recommendations by Tier

Table 11. Recommendations by tier, with evidence base, responsible parties, evidence strength, and limits.

Tier	Recommendation	Evidence base	Responsible party	Strength	Limits
Pedagogical	Teach listening as a strategic process (planning, monitoring, evaluation)	Mohammad & Alsharf (2025, p. 2148); Zenbil (2021); Bianku & Mahmud (2026)	Teaching staff	Low	No Libyan trial of metacognitive instruction; MALQ measures awareness
Pedagogical	Provide systematic bottom-up training on connected-speech features	Omar (2022, p. 20); Omar & Pilus (2022)	Course designers, teachers	Low	One intervention study; same cohort as baseline; perception outcomes
Pedagogical	Integrate authentic video with structured cooperative tasks	Aldizeeri (2025, pp. 1431, 1434)	Teachers, curriculum committees	Very low	n = 10, non-randomised; promising, not proven

Pedagogical	Build lexical depth for listening (spoken vocabulary, idioms, jargon)	Four independent samples (Table 9, Theme 2)	Teachers, materials writers	Low	Self-report only; no measured vocabulary-listening link
Pedagogical	Address anxiety via low-stakes practice, repeated listening, peer interaction	Omar (2022, p. 14); Othman (2025, p. 183)	Classroom teachers	Very low	No dedicated anxiety instrument; causal direction unknown
Pedagogical	Reduce reliance on mental translation; train directed attention	Bianku & Mahmud (2026)	Classroom teachers	Very low	Single-site awareness evidence
Institutional	Invest in audio technology and classroom acoustics	Mohammad & Alsharf (2025, pp. 2148, 2150)	University administrations	Very low	One quantitative study (n = 41)
Institutional	Supply authentic materials; train teachers in listening pedagogy and assessment	Zenbil (2021); Al-shareef (2022); Pathan et al. (2016, p. 19)	Administrations, faculties of education	Very low	Teacher-side and school-level data
Institutional	Guarantee delivery of the Listening Concepts course to all majors	Othman (2025, pp. 181-182)	Department heads	Very low	Single university, self-report
Assessment-policy	Include listening in formal examinations with piloted task specifications, school and university levels	Alghanimi & Ghazalla (2021, Abstract section)	Ministry; examination committees	Low	School-level observational evidence; effect not experimentally tested in Libya
Assessment-policy	Clarify objectives of every listening task; develop standardised tests	Al-shareef (2022, Abstract section); gap (e) in Table 12	Teachers; examination committees	Very low	Single university; no standardised Libyan instrument
Research	Test metacognitive instruction, connected-speech training, and technology-enhanced listening experimentally	Gaps (Table 12)	Researchers, graduate programmes	Not applicable	Design standards stated in Section 6.8

7.2 Future Research Agenda

Table 12. Research gaps and the studies required to close them.

Gap	Required study	Indicator to be measured
No objective listening tests	Construct and validate standardised listening tests for Libyan universities	Validity and reliability coefficients; score distributions
Small samples	Multi-university studies with adequate statistical power	Institution-level differences; effect sizes with CIs
Limited geographic representation	Samples from southern and private institutions	Between-region comparison of difficulty profiles
No longitudinal data	Track listening development across years of study	Within-student change in measured performance
Weak experimental evidence	Randomised or well-controlled quasi-experiments with comparison groups	Effect sizes with CIs; allocation integrity
Weak anxiety measurement	Studies using dedicated listening-anxiety instruments	Listening-anxiety scores and relation to performance
Limited lexical evidence	Studies linking measured vocabulary size to listening performance	Vocabulary and listening scores in the same sample
No environmental measurement	Field studies of noise, equipment, and classroom conditions	Measured acoustics and listening performance
Few qualitative studies	In-depth interviews and classroom observation	Learner experience data; observed teaching practice
No implementation research	Evaluation of actual programme delivery	Delivered versus intended curriculum; hours taught; assessment practice

8. Limitations of the Review

The limitations are organised in four levels, from the search to the conclusions; each is stated with its scientific impact.

8.1 Search and Retrieval

Scopus and Web of Science could not be searched for lack of institutional access, so indexing coverage is reduced and some eligible studies may have been missed. The twenty-result cap per string means the search was restricted rather than exhaustive; records beyond the cap were not screened. The corpus is dominated by English-language publications, so Arabic-only scholarship is likely under-represented, and one Arabic-language study is represented through its published bilingual abstract. The search window ends in November 2026, and publications from the final weeks may not have been indexed at the last search date. The evidence base cannot be claimed complete; the direction of the bias is towards omission.

8.2 Screening and Appraisal

All screening, eligibility, extraction, and appraisal decisions were originally made by both researchers. During the revision stage, a retrospective criterion-based third assessment by a designated academic reviewer (Appendix G) audited all 34 full-text eligibility decisions, all 65 MMAT criterion judgements, and all category assignments, and concurred with every one (kappa = 1.00 for the eligibility decisions; 100% per-criterion agreement, reported

descriptively without a pooled kappa, because criteria within a study are not independent); seven extracted values from Omar and Pilus (2022) were verified against the source with no corrections. This audit strengthens the consistency and auditability of the documented decisions, but it was non-blinded and retrospective: it verifies that each decision follows from its recorded justification and the criteria, and it does not provide the judgement of a third human reader screening the full texts blind, so it does not convert the review into a dual-screened one. The protocol was not prospectively registered, so deviations from the original plan cannot be audited externally. The original working screening log was not retained during preparation; the entries in Appendix C were reconstructed by re-executing the documented search strings and re-screening the retrieved records, a disclosed reconstruction that cannot restore the original contemporaneous record and cannot be guaranteed to reproduce every original screening decision. One intervention study (Omar & Pilus, 2022) was initially appraised from the publisher's abstract; the full text was retrieved in the final verification pass and the study reappraised from it.

8.3 The Primary Evidence

The eligible national literature is small (15 reports, 14 studies, 12 independent samples), so the synthesis is an integrated profile of the available evidence, not a meta-analytic estimate. Two reports concern the same Sabratha cohort; the synthesis counts them once, but the intervention findings cannot be read as a replication of the baseline findings, and undetected overlap between other samples cannot be excluded. The evidence is dominated by self-report questionnaires, so common-method variance and social-desirability effects cannot be excluded, and perceptions are not equated with proficiency anywhere in the synthesis. No study used a standardised listening proficiency instrument, so the relationship between reported difficulty and measured performance is untested. The geographic distribution omits the southern universities and private institutions; the results are a profile of the available evidence, not a representative account of all Libyan universities.

8.4 Synthesis and Inference

GRADE was applied by a single rater as an explicitly declared adaptation; it lacks the safeguards of a panel-based assessment, and its health-originated conventions had to be adapted for descriptive educational outcomes. Publication bias could not be quantified in a small national literature concentrated in a few outlets, so absence of evidence for publication bias is not evidence of its absence. The thematic synthesis integrates studies that differ in design, instruments, and constructs; although direct, contextual, and intervention evidence are reported separately and no cross-type confirmation is claimed, the themes remain interpretations of a heterogeneous corpus, and every conclusion is bounded by the certainty ratings in Table 10.

9. Conclusion

This review synthesised the available evidence on listening skill challenges in Libyan higher education published between 2015 and 2026: 15 reports representing 14 studies and 12 independent samples, with direct university-listening evidence distinguished from teaching-assessment, contextual, and intervention evidence. The findings supported directly by the corpus are the recurrence of perceptual difficulty with connected speech, the centrality of lexical gaps during listening, the interdependence of the three processing phases, and the structural neglect of listening in teaching and assessment; the environmental finding, though the most severe domain where measured, rests on one sample of 41 students.

The strength of this evidence is limited: certainty is low for the perceptual finding and very low for the remaining findings, most of the direct student evidence relies on self-reported measures from institutionally concentrated samples, and no study used a standardised listening proficiency instrument or a dedicated listening-anxiety scale. The perceived-versus-measured distinction is preserved throughout: the questionnaire themes report perceived difficulty, the test-based strand reports observed performance, and the metacognitive findings report strategic awareness as measured by MALQ, not listening proficiency; no theme mixes the two types of evidence. The current evidence therefore supports a cautious description of recurring challenges rather than a national estimate or a causal explanation.

Future research should prioritise objective listening assessment, multi-institutional sampling, dedicated measurement of listening anxiety, adequately controlled intervention studies, and longitudinal tracking of listening development, so that the next synthesis can report the measured reduction of the challenges alongside their persistence.

Declarations

Authors Contributions. The authors conceptualised the review, designed and executed the search, screened all records, extracted the data, performed the quality, risk-of-bias, and certainty appraisals, and wrote and revised the manuscript.

Funding. This research received no external funding.

Conflict of Interest. The authors declare no conflict of interest.

Data Availability. All data are drawn from published sources cited in the reference list. The extraction matrix (Appendix E) and the third-assessment verification sheets (Appendix G) are available from the authors on reasonable request.

Protocol Availability. The review protocol was not prospectively registered; the dated internal protocol and its amendment log are described in Section 3.2.

Ethics Statement. The review synthesised published sources only; no new human data were collected, and ethical approval was not required.

Acknowledgements. The authors thank the third , Abdulbaset Salah (doctoral researcher, Libyan Academy for Graduate Studies) for the retrospective verification reported in Section 4.3 and Appendix G.

References

[1] Aldarasi, R. A. (2020). *Foreign language anxiety: Libyan students speaking in English* [Doctoral thesis, University of Glasgow]. Enlighten: Theses. https://doi.org/10.5525/gla.thesis.80275

[2] Aldizeeri, M. M. A. (2025). The effect of using the cooperative approach and YouTube clips in developing the listening skill of Libyan English language students. *African Journal of Advanced Studies in Humanities and Social Sciences*, *4*(4), 1431-1438. https://aaasjournals.com/index.php/ajashss/article/view/1791

[3] Alghanimi, Y. A., & Ghazalla, N. A. (2021). Curriculum intentions and implementation of teaching the listening skill in EFL classrooms in Libyan public schools [in Arabic]. *Al-Qurtas*, *13*. https://alqurtas.alandalus-libya.org.ly/ojs/index.php/qjhar/article/view/272

[4] Al-shareef, H. A. (2022). *Listening assessment strategies employed by Libyan EFL university teachers* [Master's thesis, University of Zawia]. University of Zawia DSpace. https://dspace.zu.edu.ly/handle/1/2141

[5] Anderson, J. R. (1995). *Cognitive psychology and its implications* (4th ed.). W. H. Freeman.

[6] Bahron, N. O., & Abujela, I. M. A. (2026). The listening skill: Neglected, tested, and rarely taught: The realities of teaching listening in the Libyan context. *Comprehensive Journal of Science*, *11*-2473), 41 ملحق(2483. https://doi.org/10.65405/wdr80v48

[7] Bianku, S. N., & Mahmud, G. I. (2026). The relationship between metacognitive awareness and listening comprehension difficulties among EFL learners in Libyan universities: A case study at Omar Al-Mukhtar University. *Al-imad Journal of Humanities and Applied Sciences*, *2*(1), 25-34. https://doi.org/10.64943/ajhas.2026.020103

[8] Bin Naji, S. M., Sodiq, J., & Setyaji, A. (2026). Speaking and listening difficulties among Libyan EFL students in higher education: Challenges and implications for instructional reform. *English Review: Journal of English Education*, *14*(3), 931-942. https://doi.org/10.25134/englishreview.v14i3.264

[9] Fatah, H. A., & Farj, M. A. (2024). Listening problems that facing students in the first semesters in English department in Elmarj University [in Arabic: مشاكل الاستماع التي تواجه الطلبة في المراحل الأولى بقسم اللغة الانجليزية بكلية الآداب والعلوم]. *Majallat Jami'at Tubruq li-l-Ulum al-Ijtima'iyya wa-l-Insaniyya* [مجلة جامعة طبرق للعلوم الاجتماعية والإنسانية), 15], (778-760. Full text also available: https://www.researchgate.net/publication/382869646

[10] Field, J. (2008). *Listening in the language classroom*. Cambridge University Press.

[11] Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, *28*(4), 551-570. https://doi.org/10.1016/S0346-251X(00)00021-4

[12] Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650), 924-926. https://doi.org/10.1136/bmj.39489.470347.AD

[13] Halali, A. A. S., Elmodeer, F. M., & Fares, E. K. A. (2026). Vocabulary learning anxiety and its relationship with vocabulary learning strategies among EFL Libyan undergraduate students at Misurata University. *Bani Waleed University Journal of Humanities and Applied Sciences*, *11*(1), 262-282. https://jhas-bwu.com/index.php/bwjhas/article/view/1078

[14] Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2019). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, *34*(4), 285-291. https://doi.org/10.3233/EFI-180221

[15] Mohammad, A., & Alsharf, A. (2025). Listening comprehension difficulties of EFL students at the Faculty of Languages and Translation, Azzaytuna University. *AlQalam Journal of Medical and Applied Sciences*, *8*(4), 2148-2154. https://doi.org/10.54361/ajmas.258410

[16] Omar, N. M. (2022). Libyan EFL students' listening comprehension problems in listening to connected speech. *Academic Journal of Research and Scientific Publishing*, *4*(39), 5-20. https://doi.org/10.52132/ajrsp.e.2022.39.1

[17] Omar, N. M., & Pilus, Z. (2022). Effects of explicit connected speech instruction on Libyan EFL learners' perceptions of listening comprehension problems in speech recognition, parsing and utilisation. *IIUM Journal of Educational Studies*, *10*(1), 66-89. https://doi.org/10.31436/ijes.v10i1.425

[18] Othman, F. (2025). Challenges in developing listening skills in English as a foreign language learning (EFL). *Al-Mukhtar Journal of Social Science*, *43*(1), 168-190. https://mdp.omu.edu.ly/journals/index.php/mjssc/article/view/1279

[19] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, Article n71. https://doi.org/10.1136/bmj.n71

[20] Pathan, M. M., Al Khaiyali, A. T., & Marayi, Z. E. (2016). Teaching English as a foreign language in Libyan schools: Issues and challenges. *International Journal of English and Education*, *5*(2), 19-39. https://www.researchgate.net/publication/313240240

[21] Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*, Article 45. https://doi.org/10.1186/1471-2288-8-45

[22] Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge

[23] Zenbil, D. A. A. (2021). *Factors influencing the Libyan EFL students' listening comprehension* [Master's thesis, University of Zawia]. University of Zawia DSpace. https://dspace.zu.edu.ly/handle/1/2036.

Appendix A: Verified Reference Register

The register covers every reference in the list. For each entry it records the source type, the outlet, the volume/issue and pages, the identifier (DOI or URL), whether the full text was consulted, and the verification outcome. Every in-text citation was matched to a reference entry and every reference entry was matched to at least one in-text citation; there are no orphan citations and no orphan references. Direct quotations are marked with quotation marks and carry a page or section locator; all other source material is paraphrased with author-year citations. Two citations are secondary and are introduced as such (Yagang, 1994, as cited in Othman, 2025; Hamouda, 2013, as cited in Othman, 2025), because the original sources were not obtainable; they are not listed separately in the reference list, in accordance with APA (7th edition) practice.

Reference	Type	Published in	Vol(Issue), pages	DOI / URL	Full text consulted	Verification outcome
Aldarasi (2020)	Doctoral thesis	University of Glasgow (Enlighten: Theses)	n/a	https://doi.org/10.5525/gla.thesis.80275	Yes (repository record and text)	Verified; DOI added during revision
Aldizeeri (2025)	Journal article	African Journal of Advanced Studies in Humanities and Social Sciences	4(4), 1431-1438	https://aaasjournals.com/index.php/ajashss/article/view/1791	Yes (full text)	Verified; statistics quoted against full text
Alghanimi & Ghazalla (2021)	Journal article (Arabic)	Al-Qurtas (Al Andalus University)	13	https://alqurtas.alandalus-libya.org.ly/ojs/index.php/qjhar/article/view/272	Yes (journal page and text)	Verified
Al-shareef (2022)	Master's thesis	University of Zawia DSpace	n/a	https://dspace.zu.edu.ly/handle/1/2141	Yes (repository record)	Verified; URL corrected (earlier URL returned page not found)
Anderson (1995)	Book	W. H. Freeman (4th ed.)	n/a	—	Standard bibliographic record; three-phase model located at pp. 378-379 via published citations of	Reference added; previously an orphan citation

					the model (Ishii; Goh, 2000)	
Bahron & Abujela (2026)	Journal article	Comprehensive Journal of Science	11(ملحق 41), 2473-2483	https://doi.org/10.65405/wdr80v48	Yes (full text)	Verified via Crossref; volume/issue/pages added
Bianku & Mahmud (2026)	Journal article	Al-imad Journal of Humanities and Applied Sciences	2(1), 25-34	https://doi.org/10.64943/ajhas.2026.020103	Abstract, metadata, and journal record	Page range 25-34 confirmed from the journal record; findings cited from the abstract-level record only
Bin Naji et al. (2026)	Journal article	English Review: Journal of English Education	14(3), 931-942	https://doi.org/10.25134/englishreview.v14i3.264	Abstract and metadata	Used as contextual evidence only
Fatah & Farj (2024)	Journal article (Arabic)	Tobruk University Journal of Social Sciences and Humanities	(15), 760-778	Full text: https://www.researchgate.net/publication/382869646	Yes (full text via ResearchGate)	Statistics (76%, 70%, 50%) verified against full text; journal-of-record attribution corrected from ResearchGate to the Tobruk journal
Field (2008)	Book	Cambridge University Press	n/a	—	Standard bibliographic record	Verified
Goh (2000)	Journal article	System	28(4), 551-570	https://doi.org/10.1016/S0346-251X(00)00021-4	Abstract and publisher record	Added in this revision for the three-phase application and the recursive-phases property; abstract-level claims only
Guyatt et al. (2008)	Journal article	BMJ	336(7650), 924-926	https://doi.org/10.1136/bmj.39489.470347.AD	Standard bibliographic record	Added for the GRADE section
Halali et al. (2026)	Journal article	Bani Waleed University Journal of Humanities and Applied Sciences	11(1), 262-282	https://jhas-bwu.com/index.php/bwjhas/article/view/1078	Journal record and abstract	Verified; construct description corrected (vocabulary learning anxiety, not listening anxiety)
Hong et al. (2019)	Journal article	Education for Information	34(4), 285-291	https://doi.org/10.3233/EFI-180221	Standard bibliographic record	Added for the MMAT section
Mohammad & Alsharf (2025)	Journal article	AlQalam Journal of Medical and Applied Sciences	8(4), 2148-2154	https://doi.org/10.54361/ajmas.258410	Publisher page and full text	Both authors verified on the publisher page (Adel Mohammad; Awatif Alsharf, ORCID 0000-0002-6958-2171); the Crossref deposit omits the second author, but the reference as cited is correct

Omar, N. M. (2022)	Journal article	Academic Journal of Research and Scientific Publishing	4(39), 5-20	https://doi.org/10.52132/ajrsp.e.2022.39.1	Yes (full text)	Verified; DOI added; all quoted statistics checked against the full text
Omar & Pilus (2022)	Journal article	IIUM Journal of Educational Studies	10(1), 66-89	https://doi.org/10.31436/ijes.v10i1.425	Full text obtained and consulted (PDF downloaded from the journal platform, IIUM Press OJS, reached through the DOI; 25 pp.)	Added during the second revision round; full text retrieved in the final verification pass (14 November 2026). The full text confirms N = 62 Libyan learners aged 19-21 at Sabratha University, a pretest-posttest control group design self-identified as quasi-experimental, with four intact classes randomly assigned at cluster level to treatment (n = 32) and control (n = 30) (pp. 71-72) while subject selection was not randomised (p. 77), a 12-week intervention of 24 sessions, an adapted LCPQ content-validated by three experts and piloted (alpha = 0.90 recognition, 0.89 parsing, 0.89 utilisation), non-significant pre-test differences (p = 0.387, 0.311, 0.347), GEE adjusted for gender and age, effect sizes of d = 1.19 (recognition), 1.14 (parsing), and 1.15 (utilisation), and the authors' explicit statement that subject selection was not randomised (p. 77)

Othman (2025)	Journal article	Al-Mukhtar Journal of Social Science	43(1), 168-190	https://mdp.omu.edu.ly/journals/index.php/mjssc/article/view/1279	Yes (full text)	Verified; the DOI printed in the article PDF (10.54172/tbthan11) does not resolve in the DOI system, so the stable journal URL was used
Page et al. (2021)	Journal article	BMJ	372, Article n71	https://doi.org/10.1136/bmj.n71	Standard bibliographic record	Verified
Pathan et al. (2016)	Journal article	International Journal of English and Education	5(2), 19-39	https://www.researchgate.net/publication/313240240	Yes (full text)	Verified; pagination confirmed
Thomas & Harden (2008)	Journal article	BMC Medical Research Methodology	8, Article 45	https://doi.org/10.1186/1471-2288-8-45	Standard bibliographic record	Verified
Vandergrift & Goh (2012)	Book	Routledge	n/a	—	Standard bibliographic record	Verified
Zenbil (2021)	Master's thesis	University of Zawia DSpace	n/a	https://dspace.zu.edu.ly/handle/1/2036	Yes (repository record)	Verified; URL corrected (earlier URL returned page not found)

The PRISMA flow counts (Table 5) were reconstructed under the counting rules stated in Section 4.1 and reconciled for arithmetic consistency (346 identified; 58 duplicates removed before screening; 288 screened at title/abstract; 254 excluded; 34 reports sought, all retrieved; 19 excluded with reasons; 15 reports included; 14 studies; 12 independent samples); the full-text screening log is reported in Appendix C.

Appendix B: Full Search Strategy by Platform

The final search across all databases and platforms was conducted up to 14 November 2026. Strings SS-1 to SS-12 were executed in Google Scholar, ERIC, ResearchGate, and each of the Libyan platforms and repositories listed in Section 4.2; strings SS-13 and SS-14 (Arabic) were executed in Google Scholar and the Libyan platforms, since ERIC indexes English-language records only.

Counting basis: for each string, the results retrieved across all platforms were merged and de-duplicated within the string, and the first twenty ranked results of the merged list were screened. The per-string counts in Table 2 therefore report screened records (20 per string; 14 × 20 = 280), not the total number of results retrieved. Total retrieved counts were not systematically logged, because platform result lists overlap heavily and the review's counting unit is the screened record; the supplementary Crossref query (SS-15) was logged separately (20 records retrieved and screened).

ID	Search string as executed	Platform(s)	Records screened
SS-1	("listening comprehension" OR "listening skill" OR "listening skills" OR "listening problems" OR "listening difficulties") AND (Libya OR Libyan) AND (EFL OR "English language teaching" OR universit* OR "English department")	Google Scholar; ERIC; ResearchGate; Libyan platforms	20 (merged)
SS-2	("listening anxiety" OR "listening apprehension") AND (Libya OR Libyan) AND (EFL OR "English language teaching" OR universit*)	Google Scholar; ERIC; ResearchGate; Libyan platforms	20 (merged)
SS-3	("listening strategies" OR "listening strategy instruction" OR "listening strategy use") AND (Libya OR Libyan) AND (EFL OR universit*)	Google Scholar; ERIC; ResearchGate; Libyan platforms	20 (merged)
SS-4	("listening assessment" OR "assessing listening" OR "listening test*") AND (Libya OR Libyan) AND (EFL OR universit*)	Google Scholar; ERIC; ResearchGate; Libyan platforms	20 (merged)
SS-5	("metacognitive awareness listening" OR "Metacognitive Awareness Listening Questionnaire" OR MALQ) AND (Libya OR Libyan)	Google Scholar; ERIC; ResearchGate; Libyan platforms	20 (merged)

SS-6	("listening comprehension problems" OR "listening difficulties") AND (Libya OR Libyan) AND ("higher education")	Google Scholar; ERIC; ResearchGate; Libyan platforms	20 (merged)
SS-7	("listening skill" OR "listening skills") AND (Libya OR Libyan) AND ("faculty of arts")	Google Scholar; ERIC; ResearchGate; Libyan platforms	20 (merged)
SS-8	("listening comprehension" OR "listening difficulties") AND (Libya OR Libyan) AND ("English majors")	Google Scholar; ERIC; ResearchGate; Libyan platforms	20 (merged)
SS-9	("teaching listening" OR "listening instruction") AND (Libya OR Libyan)	Google Scholar; ERIC; ResearchGate; Libyan platforms	20 (merged)
SS-10	("connected speech") AND (Libya OR Libyan) AND (listening)	Google Scholar; ERIC; ResearchGate; Libyan platforms	20 (merged)
SS-11	("listening comprehension") AND ("Libyan universities")	Google Scholar; ERIC; ResearchGate; Libyan platforms	20 (merged)
SS-12	(listening) AND ("Libyan EFL" OR "Libyan EFL learners")	Google Scholar; ERIC; ResearchGate; Libyan platforms	20 (merged)
SS-13	(مهارة الاستماع OR الاستماع) AND (ليبيا OR الليبية OR الجامعات) AND (أقسام اللغة الإنجليزية OR طلبة الجامعة)	Google Scholar; Libyan platforms	20 (merged)
SS-14	("تدريس الاستماع" OR "تعليم الاستماع" OR "صعوبات الاستماع") AND (ليبيا) AND (اللغة الإنجليزية)	Google Scholar; Libyan platforms	20 (merged)
SS-15	"listening difficulties Libyan EFL university students" (Crossref metadata search, filter 2015-2026)	Crossref	20
Total screened at title/abstract level from string searching			**280 + 20 (Crossref)**

Records were also identified through backward citation chaining (reference lists of retrieved studies) and forward citation chaining (works citing them); 46 unique records not already captured by the string searches were added this way and merged with the string-search results before duplicate removal. Google Scholar results beyond the first twenty per merged string were inspected opportunistically but were not counted in the flow numbers in Table 5.

Note (Table 2): two verification searches were executed on 14 November 2026 (a Crossref bibliographic query restricted to 2024-2026 and an ERIC query combining Libya, listening, and higher education terms) to check for missed eligible records; neither identified an additional eligible record, so neither appears as a counted row in Table 2 or in the flow counts in Table 5 (Section 4.1).

Appendix C: Full-Text Screening Log

Thirty-four reports were assessed in full text. The table below reports the disposition of every record with full bibliographic data. Note on the provenance of the excluded records: the original working screening log was not retained during preparation of the revised manuscript, so the nineteen excluded records below were reconstructed by re-executing the documented search strings on the same platforms and re-screening the retrieved records against the same criteria. Each entry is a real publication whose existence and bibliographic details were verified against Crossref, the publisher's site, or the document itself, and the exclusion reason records the re-screening decision. Verification level: most excluded records were verified against the full text or the publisher or Crossref record at the time of reconstruction; for two national-journal entries the DOI is registered but the journal server was unreachable at verification, so verification rests on the Crossref record; no excluded record proved unretrievable. Because the original contemporaneous log was not retained, the reconstruction cannot be guaranteed to reproduce every original screening decision; it is reported as a documented, verified approximation rather than as the original record. The fifteen included sources (FT-20 to FT-34) were never part of the affected material, and their dispositions are documented directly.

Record ID	Author(s)	Year	Title	Journal / University	Found via	Decision	Reason
FT-01	Saleh, S. E.	2017	Some Libya EFL University Students' Attitudes towards Using	Theory and Practice of Second Language Acquisition	Google Scholar	Excluded	Other skill only (reading; no listening component)

			Authentic Materials for Reading Classes				
FT-02	Ramadan, M. O.	2019	The Use of Connectors in the Writing of Libyan EFL Students: A Study in the English Department, Sirte University	Journal of Academic Research (Sirte), 13	Google Scholar	Excluded	Other skill only (writing; no listening component)
FT-03	Elshawish, M. F.	2023	Speaking Anxiety among English Majors at Misurata University: A Learner's Perspective	Journal of Academic Research (Misurata), 27, 11-25	Google Scholar	Excluded	Other skill only (speaking anxiety; no listening measure)
FT-04	Aldbea, K. A. H.	2024	The Difficulties That Faced Libyan Students in Learning English Vocabulary and Understanding Academic Texts	International Journal for Multidisciplinary Research, 6(6)	Google Scholar	Excluded	Other skill only (vocabulary; no listening outcome)
FT-05	Haroun, A., & AL-Darraji, O.	2022	Difficulties Facing Libyan Universities Students in Learning English Language Conversations	Libyan International Journal	ResearchGate	Excluded	Other skill only (conversation; no listening outcome)
FT-06	Rafah, A. A., & Elraggas, A. A.	2024	The Difficulties of Translating English Phrasal Verbs into Arabic: The Case of Libyan EFL Graduate Translation Students at the Libyan Academy	Afāq Journal for Humanities and Applied Studies, 2, 428-454	Crossref	Excluded	Other skill only (translation; no listening component)
FT-07	Alkhelaiwi, S.	2023	Are You Listening Strategically	World Journal of English	ERIC	Excluded	Other country

			or Randomly? Examining Listening Comprehension Strategies among Saudi University EFL Students	Language, 13(5)			(non-Libyan context)
FT-08	Borekci, R., & Yavuz, F.	2018	Foreign Language Listening Anxiety among Turkish EFL Learners	Global Journal of Foreign Language Teaching, 7(2)	Google Scholar	Excluded	Other country (non-Libyan context)
FT-09	Shahani, S., Chalak, A., & Tabrizi, H. H.	2022	Impact of Critical Thinking Instruction through Flipped Teaching on Iranian EFL Learners' Listening Comprehension	Turkish Online Journal of Distance Education	ERIC	Excluded	Other country (non-Libyan context)
FT-10	Ali, A., & Hameed, D. T.	2020	The Relationship between Listening Anxiety among Iraqi EFL University Students and Their Performance in Listening Comprehension	Kirkuk University Journal for Humanities Studies, 15(1)	ERIC	Excluded	Other country (non-Libyan context)
FT-11	Agustiana, V.	2019	Listening Anxiety among Indonesian EFL Students	Indonesian EFL Journal, 5(1)	Crossref	Excluded	Other country (non-Libyan context)
FT-12	Omar, M. A.	2018	Investigating Strategies Used by Libyan EFL Teachers in Teaching Reading Comprehension at Sirte Basic Education Schools	Sirte University research journal, 12	Libyan platforms	Excluded	Pre-university context without direct relevance to university preparation
FT-13	Altaieb, S.	2015	Obstacles Libyan	Journal of Modern	Google Scholar	Excluded	Pre-university

			Teachers of English Encounter While Implementing English Language Curriculum in Libyan High Schools	Education Review, 5(9)			context without direct relevance to university preparation
FT-14	Orafi, S., & Idris, H.	2023	Teachers' Perceptions of Teaching English Language to Young Learners in Libyan Primary Schools in Al-Marj Region	African Journal of Advanced Studies in Humanities and Social Sciences, 2(4), 610-619	Google Scholar	Excluded	Pre-university context without direct relevance to university preparation
FT-15	Libya Herald (news report)	2025	UK Minister Visits Tripoli: Agreement Signed to Improve Quality of English Teaching at Libyan Educational Establishments	libyaherald.com	Google Scholar	Excluded	Non-scholarly outlet
FT-16	Libya Observer (news brief)	2023	Misrata Launches English Teaching Program	libyaobserver.ly	Google Scholar	Excluded	Non-scholarly outlet
FT-17	Mustapha, S. M., & Mahmoud, F. T. M.	2019	Autonomous Learning Readiness and English Language Performance of EFL Libyan Secondary School Students	Proceedings of the Second International Conference on Social, Economy, and Education, 109-116	ResearchGate	Excluded	Population not Libyan university learners (secondary school)
FT-18	Algamati, I. F.	2024	Speaking Difficulties Encountered by Libyan Secondary School Students	Scientific Journal of Faculty of Education, Misurata University, 9(24)	Libyan platforms	Excluded	Population not Libyan university learners (secondary school)
FT-19	Fatah, H. A., & Farj, M. A.	2024	Listening Problems That Facing	ResearchGate	ResearchGate	Excluded	Duplicate report of data already

	(ResearchGate record)		Students in the First Semesters in English Department in Elmarj University (ResearchGate record of the journal article)				represented in the corpus (journal version, FT-27)
FT-20	Pathan, M. M., Al Khaiyali, A. T., & Marayi, Z. E.	2016	Teaching English as a foreign language in Libyan schools: Issues and challenges	International Journal of English and Education, 5(2), 19-39	ERIC	Included	—
FT-21	Aldarasi, R. A.	2020	Foreign language anxiety: Libyan students speaking in English	University of Glasgow (doctoral thesis)	Citation chaining	Included	—
FT-22	Alghanimi, Y. A., & Ghazalla, N. A.	2021	Curriculum intentions and implementation of teaching the listening skill in EFL classrooms in Libyan public schools	Al-Qurtas, 13	Libyan platforms	Included	—
FT-23	Zenbil, D. A. A.	2021	Factors influencing the Libyan EFL students' listening comprehension	University of Zawia (master's thesis)	Libyan platforms	Included	—
FT-24	Omar, N. M.	2022	Libyan EFL students' listening comprehension problems in listening to connected speech	Academic Journal of Research and Scientific Publishing, 4(39), 5-20	Google Scholar	Included	—
FT-25	Omar, N. M., & Pilus, Z.	2022	Effects of explicit connected speech instruction on Libyan EFL learners' perceptions of listening comprehensio	IIUM Journal of Educational Studies, 10(1), 66-89	Citation chaining (forward, from Omar, 2022)	Included	Same cohort as FT-24; not an independent sample

			n problems in speech recognition, parsing and utilisation				
FT-26	Al-shareef, H. A.	2022	Listening assessment strategies employed by Libyan EFL university teachers	University of Zawia (master's thesis)	Libyan platforms	Included	—
FT-27	Fatah, H. A., & Farj, M. A.	2024	Listening problems that facing students in the first semesters in English department in Elmarj University	Tobruk University Journal of Social Sciences and Humanities, (15), 760-778	Google Scholar	Included	—
FT-28	Othman, F.	2025	Challenges in developing listening skills in English as a foreign language learning (EFL)	Al-Mukhtar Journal of Social Science, 43(1), 168-190	Libyan platforms	Included	—
FT-29	Mohammad, A., & Alsharf, A.	2025	Listening comprehension difficulties of EFL students at the Faculty of Languages and Translation, Azzaytuna University	AlQalam Journal of Medical and Applied Sciences, 8(4), 2148-2154	Libyan platforms	Included	—
FT-30	Aldizeeri, M. M. A.	2025	The effect of using the cooperative approach and YouTube clips in developing the listening skill of Libyan English language students	African Journal of Advanced Studies in Humanities and Social Sciences, 4(4), 1431-1438	Google Scholar	Included	—
FT-31	Halali, A. A. S., Elmodeer, F. M., & Fares, E. K. A.	2026	Vocabulary learning anxiety and its relationship	Bani Waleed University Journal of Humanities and Applied	Citation chaining	Included	Contextual evidence (vocabulary-anxiety construct)

			with vocabulary learning strategies among EFL Libyan undergraduate students at Misurata University	Sciences, 11(1), 262-282			
FT-32	Bianku, S. N., & Mahmud, G. I.	2026	The relationship between metacognitive awareness and listening comprehension difficulties among EFL learners in Libyan universities: A case study at Omar Al-Mukhtar University	Al-imad Journal of Humanities and Applied Sciences, 2(1), 25-34	Crossref (supplementary search)	Included	—
FT-33	Bin Naji, S. M., Sodiq, J., & Setyaji, A.	2026	Speaking and listening difficulties among Libyan EFL students in higher education: Challenges and implications for instructional reform	English Review: Journal of English Education, 14(3), 931-942	Crossref (supplementary search)	Included	Contextual evidence (narrative review)
FT-34	Bahron, N. O., & Abujela, I. M. A.	2026	The listening skill: Neglected, tested, and rarely taught: The realities of teaching listening in the Libyan context	Comprehensive Journal of Science, 11(41 ملحق), 2473-2483	Citation chaining	Included	Contextual evidence (conceptual article)

Appendix D: MMAT Appraisal Records (Original 2018 Categories)

Each study was first classified by its actual design as reported in its methods section, then rated on the two screening questions and the five criteria of its own MMAT 2018 category. Judgements: Y = yes; N = no; CT = cannot tell from the consulted record. No aggregate score was computed. Justifications are tied to what the consulted record reports.

Screening questions (all empirical studies). SQ1 (Are the research questions clear?): Y for all 13 empirical studies; each states a clear objective or question. SQ2 (Do the collected data allow the research questions to be addressed?): Y for all 13; each collected data of the type its question requires. A uniform rule governs criterion 4.4 (nonresponse bias) across the survey studies: Y is assigned only where the source reports a response rate or

documents full participation; where only the administration procedure is described, the criterion is CT, because described administration alone does not establish low nonresponse bias.

Study; category	C1: judgement and justification	C2	C3	C4	C5
Pathan et al. (2016); Quantitative descriptive	4.1 Y: teacher survey relevant to the question of school ELT conditions	4.2 CT: representativeness of the 12 schools not established	4.3 Y: questionnaire data reported completely for the items used	4.4 CT: response rate and nonresponse bias not reported	4.5 Y: descriptive analysis appropriate
Aldarasi (2020); Qualitative	1.1 Y: qualitative interview approach appropriate to an anxiety experience question	1.2 Y: semi-structured interviews with students and expert teachers adequate	1.3 Y: findings grounded in reported extracts	1.4 Y: interpretation substantiated by data	1.5 CT: coherence across sources, collection, analysis, and interpretation only partially documented
Alghanimi & Ghazalla (2021); Quantitative descriptive	4.1 Y: sampling of teachers and classrooms relevant to implementation question	4.2 CT: representativeness of Tripoli classrooms not established	4.3 Y: questionnaire and observation measure implementation directly	4.4 CT: response handling not reported	4.5 Y: analysis appropriate
Zenbil (2021); Quantitative descriptive (test and questionnaire strands analysed separately, not integrated)	4.1 Y: first-year students and teachers relevant to the factors question	4.2 CT: single site, intact groups; representativeness not established	4.3 Y: test plus questionnaire appropriate	4.4 CT: nonresponse not reported	4.5 Y: analysis appropriate
Omar (2022); Quantitative descriptive	4.1 Y: purposive sample of majors relevant to the processing-problems question	4.2 CT: single site, 90% female; target population not defined	4.3 Y: validated 38-item instrument (alpha = 0.924)	4.4 CT: administration procedure reported, but response rate and respondent-nonrespondent differences not reported, so low nonresponse bias cannot be established	4.5 Y: descriptive and correlational statistics appropriate
Omar & Pilus (2022); Randomised (cluster allocation; pretest-posttest control group self-identified as quasi-experimental; four intact classes randomly assigned at cluster level, pp. 71-72; subject selection not randomised, p. 77)	2.1 CT: random assignment performed and stated twice (pp. 71-72), but the randomisation procedure is not described, only four intact classes were allocated, and the authors frame the design as quasi-experimental, so appropriate randomisation cannot be verified	2.2 Y: pre-test equivalence tested with no significant differences (p = 0.387, 0.311, 0.347; p. 78) and GEE adjusted for gender and age (p. 77)	2.3 N: cases lacking pre-test data eliminated from the analysis and the final analysed n not reported; the authors acknowledge the reduction affected the sample size (p. 85)	2.4 CT: no blinding described (pp. 76-77); outcomes were self-report questionnaires, which limits the meaning of assessor blinding	2.5 CT: 24 sessions over 12 weeks described in detail within regular classes, but no attendance or fidelity check reported (pp. 71, 76-77)
Al-shareef (2022); Mixed methods	5.1 Y: rationale for combining questionnaire and interviews stated	5.2 Y: components address the same question	5.3 Y: outputs interpreted jointly	5.4 N: divergences between strands not addressed	5.5 Y: each strand follows its own tradition
Fatah & Farj (2024); Mixed methods	5.1 Y: rationale for questionnaire, interviews, and tasks stated	5.2 CT: integration of the strands described briefly	5.3 Y: outputs interpreted	5.4 N: divergences not addressed	5.5 CT: psychometric and audit-trail documentation insufficient to judge
Othman (2025); Quantitative descriptive	4.1 Y: undergraduates relevant to the challenges question	4.2 CT: single site; representativeness not established	4.3 CT: instrument validity and reliability not fully documented	4.4 CT: administration procedure reported, but response rate and nonresponse	4.5 Y: analysis appropriate

				handling not reported	
Mohammad & Alsharf (2025); Quantitative descriptive	4.1 Y: undergraduates in semesters 4-8 relevant	4.2 CT: single site; representativeness not established	4.3 Y: 22-item instrument with reported reliability (alpha = 0.881)	4.4 CT: administration procedure reported, but response rate and nonresponse handling not reported	4.5 Y: SPSS analysis transparently reported
Aldizeeri (2025); Non-randomised	3.1 N: ten students cannot represent the target population	3.2 Y: pre-post comprehension scores appropriate to the outcome	3.3 Y: outcome data reported for both groups	3.4 N: confounders not accounted for in design or analysis	3.5 Y: intervention delivered as described
Halali et al. (2026); Quantitative descriptive	4.1 CT: sampling strategy not fully described in the consulted record	4.2 CT: representativeness not established	4.3 CT: psychometric documentation partial	4.4 CT: nonresponse not reported	4.5 Y: SPSS analysis appropriate
Bianku & Mahmud (2026); Quantitative descriptive	4.1 Y: majors with MALQ relevant to the awareness question	4.2 N: purposive single-site sample; not representative of the target population	4.3 Y: validated MALQ	4.4 CT: response handling not reported	4.5 Y: correlational analysis appropriate to the question

Sources not appraisable with MMAT: Bahron & Abujela (2026) is a conceptual article and Bin Naji et al. (2026) is a narrative review; the MMAT is designed for empirical studies only. Both were appraised narratively: Bahron & Abujela presents a clear argument with no systematic method; Bin Naji et al. reports a structured search across major databases with a reflexive thematic analysis, but its synthesis is narrative (non-PRISMA), reports no effect sizes, and may overlap with primary studies in this corpus. Neither contributes to any prevalence claim in this review.

How the appraisal was used: the criterion-level weaknesses above are the primary basis for the risk-of-bias judgements in Table 8 and the GRADE downgrades in Table 10; the studies with the weakest documentation (Fatah & Farj, 2024; Aldizeeri, 2025) are given interpretive rather than confirmatory weight in the synthesis; and no study was excluded on quality grounds, because exclusion would have removed most of a small national evidence base.

Appendix E: Data Extraction and Evidence Matrix

Extraction rule applied to every field: data not reported in the source were never recorded as facts, and no characteristic of a sample or instrument was inferred from a title or abstract; fields that could not be located in the consulted record are marked "not reported".

Study	Population and sample	Design	Instrument	Psychometrics	Key results (with locators)	Intervention	Study-reported or observed limits
Omar (2022)	62 third-year English majors, Sabratha University (90% female)	Quantitative descriptive	LCPQ, 38 items	alpha = 0.924	Overall M = 3.20 (SD = 0.75); perception M = 3.14, parsing M = 3.25, utilisation M = 3.21 (p. 13); speed M = 3.56, lexical density M = 3.55, connected-speech features M = 3.45, anxiety item M = 3.94 (p. 14); parsing items 3.63-3.79 (p. 16); idiom texts M = 3.60 (p. 17); r = .789-.842 among phases (p. 19)	None	Single site; self-report only
Omar & Pilus (2022)	62 learners aged 19-21, Sabratha University (same cohort; four intact classes)	Pretest-posttest control group, self-identified as quasi-experimental; four intact classes randomly assigned at cluster level (treatment n =	Adapted LCPQ, pre/post	alpha = 0.90 (recognition), 0.89 (parsing), 0.89 (utilisation)	Pre-test differences not significant (p = 0.387, 0.311, 0.347; p. 78); significant group x time interaction for all three categories (Wald chi-square 14.472-16.676, p < 0.001); d = 1.19 (recognition), 1.14 (parsing), 1.15	Explicit connected-speech instruction, 12 weeks, 24 sessions	Cluster-level random allocation of four classes with an undescribed randomisation procedure; selection not randomised (p. 77); no blinding or fidelity check; cases

Study	Population and sample	Design	Instrument	Psychometrics	Key results (with locators)	Intervention	Study-reported or observed limits
		32, control n = 30; pp. 71-72); subject selection not randomised (p. 77)			(utilisation); control d = 0.01-0.02 (pp. 80-83)		without pre-test data eliminated and final analysed n not reported (p. 85); perception-based outcomes; no confidence intervals
Zenbil (2021)	40 first-year students, 30 teachers, University of Zawia	Mixed methods	Listening test; questionnaire	Not reported	Students "lack some cognitive skills and strategies such as retrieving and recognition and substitution strategy" (Abstract); teachers report materials, equipment, and training deficits (Abstract)	None	Single site; test psychometrics unreported
Al-shareef (2022)	Teachers of listening courses, University of Zawia (6 interviewed)	Mixed methods	Questionnaire; semi-structured interviews	Not reported	Goals set but objectives rarely clarified; weaknesses in scoring, task setting, text selection, experience, resources, training (Abstract)	None	Teacher-side only; questionnaire psychometrics unreported
Fatah & Farj (2024)	30 first-semester majors, University of Benghazi, Almarj branch	Mixed methods	Questionnaire; interviews; in-class tasks	Not reported	76% limited vocabulary; 70% no listening practice; 50% daily input (Abstract)	None	Small sample; no psychometric documentation
Othman (2025)	56 undergraduates, Ajdabiya University (89.3% female)	Quantitative descriptive	Five-domain questionnaire	Partial	55.4% never studied the Listening Concepts course (pp. 181-182); 37.5% misrecognise known words; 41.1% familiar pronunciation without recognition (p. 183); 48.2% fear not understanding; 46.4% stop on unknown words; 42.9% cannot infer in real time (p. 183); academic performance affected (p. 168)	None	Single site; self-report
Mohammad & Alsharf (2025)	41 undergraduates, semesters 4-8, Azzaytuna University	Quantitative descriptive	22-item Likert questionnaire	alpha = 0.881 (sections .875-.886)	Physical setting M = 4.30 (SD = 0.91); listener-related M = 3.60; material-related M = 3.38; overall M = 3.76 (SD = 1.07); limited vocabulary M = 4.39 (SD = 0.92) (pp. 2148, 2150)	None	Single site; self-report
Aldizeeri (2025)	10 students, Al-Marqab University	Quasi-experimental	Pre-post listening scores; questionnaire; thematic analysis	Not reported	Significant comprehension and motivation gains in the cooperative condition; no significant gain in	Cooperative learning with YouTube clips	n = 10; no randomisation; confounding uncontrolled

Study	Population and sample	Design	Instrument	Psychometrics	Key results (with locators)	Intervention	Study-reported or observed limits
					control (pp. 1431, 1433-1434)		
Halali et al. (2026)	97 undergraduates, Misurata University	Quantitative descriptive	Online questionnaire	Partial	Moderate vocabulary anxiety M = 2.7951, particularly during listening and speaking (Abstract)	None	Construct adjacent to listening; sampling not fully described
Bianku & Mahmud (2026)	48 third- and fourth-year majors, Omar Al-Mukhtar University	Quantitative descriptive (cross-sectional survey)	MALQ, 21 items	Validated instrument	Problem-Solving M = 5.11/6; Directed Attention M = 4.45; Planning-Evaluation M = 4.46; Mental Translation M = 5.06 (Abstract)	None	Single site; awareness measure, not proficiency
Pathan et al. (2016)	35 EFL teachers, 12 schools	Survey	Teacher questionnaire	Not reported	Insufficient training, weak coordination, infrastructure shortages (p. 19)	None	School level; no listening-specific measure
Alghanimi & Ghazalla (2021)	EFL teachers and classrooms, Tripoli public schools	Quantitative	Questionnaire; observation	Not reported	Listening materials "ignored and skipped"; washback from examination exclusion (Abstract)	None	Single city; school level
Aldarasi (2020)	22 students, 8 expert teachers (doctoral research)	Qualitative	Semi-structured interviews	Not applicable	Anxiety bound up with accents, settings, social comparison (Abstract)	None	Speaking-centred construct
Bin Naji et al. (2026)	50 peer-reviewed articles (2022-2026)	Narrative review	Reflexive thematic analysis	n/a	Teacher-centred pedagogy, exam-oriented curricula, anxiety, memory overload, weak self-regulation (Abstract)	None	Narrative synthesis; possible overlap with primary studies
Bahron & Abujela (2026)	Not applicable	Conceptual	Argumentative analysis	n/a	Listening overlooked, tested rather than taught; outdated methods counterproductive (Abstract)	None	No empirical data

Report-to-sample linkage: the two Sabratha reports (Omar, 2022; Omar & Pilus, 2022) share one independent sample of 62 learners and are counted once in all prevalence-relevant counts; the intervention report is never treated as confirmation of the baseline report. No other overlap was identifiable from the available records.

Appendix F: Linkage of Discussion Conclusions to Supporting Evidence

Every substantive conclusion in Section 6 is listed below with the study or studies that support it, the evidence type it derives from, and its level of support: directly supported (the finding is reported by the cited study), theory-consistent (the finding aligns with established models or international evidence), or review interpretation (a plausible reading proposed by this review, which the data do not establish).

Discussion conclusion (section)	Supporting study or studies	Evidence type	Level of support
Perceptual difficulty with connected speech is the most consistently documented challenge (6.2)	Omar (2022); Zenbil (2021, test strand); Othman (2025); Mohammad & Alsharf (2025)	Direct university-listening	Directly supported, in perceptual and test-based strands
Print-based vocabulary learning contributes to recognition failure (6.2)	Mohammad & Alsharf (2025); Omar (2022)	Direct (self-report)	Theory-consistent; exposure, vocabulary size, and decoding never measured jointly
Limited exposure constrains the automatism of decoding (6.2)	Fatah & Farj (2024) (50% daily input)	Direct (self-report)	Theory-consistent
School-level examination washback suppresses listening instruction (6.2, 6.6)	Alghanimi & Ghazalla (2021)	Contextual (school)	Directly supported at school level; review interpretation at university level

Explicit decoding training reduces reported processing problems (6.2, 6.8)	Omar & Pilus (2022)	Intervention (perception outcomes)	Directly supported for reported problems; unproven for measured proficiency
Lexical gaps are central to listening difficulty (6.3)	Fatah & Farj (2024); Omar (2022); Othman (2025); Mohammad & Alsharf (2025)	Direct (self-report)	Directly supported as reported difficulty; vocabulary size never measured
Vocabulary weakness causes low listening performance (6.3)	None	Not applicable	Review interpretation; untested in this corpus
The three processing phases are interdependent (6.4)	Omar (2022, r = .789-.842)	Direct (correlational)	Directly supported as association; causality not established
Directed attention and planning-evaluation are the weakest strategy areas; mental translation is heavy (6.4)	Bianku & Mahmud (2026)	Direct (MALQ awareness)	Directly supported for strategy awareness, not proficiency
Strategy training would improve listening performance (6.4)	None	Not applicable	Review interpretation; hypothesis for intervention research
Anxiety is a prominent barrier during listening (6.5)	Omar (2022, M = 3.94); Othman (2025, 48.2%)	Direct (self-report items)	Directly supported as reported fear; no dedicated listening-anxiety instrument used
Anxiety causes poor listening performance (6.5)	None	Not applicable	Review interpretation; direction untested
Affect responds to pedagogy (6.5, 6.8)	Aldizeeri (2025)	Intervention (n = 10)	Directly supported within the study; not generalisable
Teachers set goals but rarely clarify objectives; assessment weaknesses are documented (6.6)	Al-shareef (2022); Zenbil (2021)	Direct teaching-assessment	Directly supported, teacher-side data
A majority of one sample never took the dedicated listening course (6.6)	Othman (2025)	Direct (self-report)	Directly supported in one sample
Examination policy is a promising lever for reform (6.6, 6.10)	Alghanimi & Ghazalla (2021); Al-shareef (2022)	Contextual + teaching-assessment	Theory-consistent; causal effect untested at university level
Physical-setting difficulty is severe where measured (6.7)	Mohammad & Alsharf (2025)	Direct (one sample, n = 41)	Directly supported in one sample; no national ranking follows
Listening outcomes respond to instructional design (6.8)	Omar & Pilus (2022); Aldizeeri (2025)	Intervention	Theory-consistent; effectiveness unproven in both studies
The Libyan profile parallels international EFL patterns (6.9)	Goh (2000); Hamouda (2013, as cited in Othman, 2025)	International comparison	Theory-consistent; not an equivalence claim
The institutional layer belongs in the analytical framework (6.10)	Alghanimi & Ghazalla (2021); Al-shareef (2022); Othman (2025); Mohammad & Alsharf (2025)	Contextual + direct	Review interpretation (organisational synthesis), not a tested causal finding

Appendix G: Retrospective Third-Assessment Verification Record

During the manuscript revision stage, a criterion-based sthird assessment of the documented review decisions was conducted by a designated third assessor specialized in accounting; statistical reviewing. The assessment audited each decision against the eligibility criteria (Table 3), the recorded page-anchored justifications (Appendices C and D), and, where a full text was available to the assessor, the source document itself. It was a verification of documented decisions, not a blinded re-screening; the agreement statistics below therefore describe the criterion-consistency of the audit trail and are expected to overstate the agreement that blinded dual screening would produce.

Unit of analysis and statistics. The two decision types are reported separately and never pooled into a single coefficient. For the eligibility audit, the unit of analysis is the full-text report decision, a binary Included/Excluded judgement applied once per report (n = 34); Cohen's kappa is computable and meaningful for this set because both assessors used both categories and each report contributes exactly one paired decision. For the MMAT audit, the 65 criterion judgements are not statistically independent, because the five criteria within each study share the same study, the same record, and often the same reported information; a single pooled kappa across them would overstate the effective sample size and is therefore not computed. Agreement is reported descriptively per criterion instead. The title-and-abstract screening stage was not re-assessed, and no agreement statistic is reported for it.

Verification scope	Units assessed	Result	Agreement statistic	Disagreements	Effect on final decisions
Full-text eligibility decisions (Appendix C)	34 report-level decisions (19 excluded, 15 included)	All 34 concurred	Raw agreement 100% (34/34); Cohen's kappa = 1.00 (unit: one binary decision per report)	0	None; PRISMA counts unchanged
MMAT category assignments (Appendix D)	13 studies	All 13 concurred	Descriptive agreement 13/13	0	None

Verification scope	Units assessed	Result	Agreement statistic	Disagreements	Effect on final decisions
MMAT criterion-level judgements (Appendix D)	65 judgements (13 studies x 5 criteria)	All 65 concurred	Descriptive agreement 100% on each criterion (13/13 paired judgements per criterion); no pooled kappa, because criteria within a study are not statistically independent	0	None
Extracted values, Omar & Pilus (2022) (Appendix E)	7 values checked against the retrieved full text	7 confirmed, 0 corrected	100% confirmation	0	None
GRADE arithmetic consistency (Table 10)	8 outcome-level ratings	All 8 final levels follow arithmetically from the starting level and the recorded downgrades	8/8 consistent	0	None

Interpretation rules applied in the audit: the uniform criterion-4.4 rule (Yes only where a response rate or full participation is documented; otherwise Cannot tell) was applied as stated in Appendix D; the Omar & Pilus category and 2.x judgements were checked against the retrieved full text (random assignment at cluster level, pp. 71-72; non-random subject selection, p. 77; pre-test equivalence, p. 78; eliminated cases, p. 85); and Cohen's kappa was computed for the eligibility decisions alone, from the paired report-level decisions, with the standard large-sample formula and the Landis-Koch band as a guide only.

Limitations of the verification: the third assessment was retrospective and non-blinded; the assessor had access to the primary decisions and their justifications; the assessor's specialisation is in accounting and research methodology rather than applied linguistics, so his role was confined to statistical and procedural verification; and the assessment does not constitute blinded dual screening of the full texts. It verifies the internal consistency and criterion-grounding of every documented decision; the single-researcher limitation stated in Section 8 is retained in full.

Appendix H: Intervention Study Details

Table 13. Characteristics and results of the intervention studies.

Element	Omar & Pilus (2022)	Aldizeeri (2025)
University	Sabratha University	Al-Marqab University
Sample	62 learners aged 19-21 (same cohort as Omar, 2022)	10 Libyan EFL students
Design	Pretest-posttest control group, self-identified as quasi-experimental; four intact classes randomly assigned at cluster level to treatment (n = 32) and control (n = 30) (pp. 71-72); subject selection not randomised (p. 77)	Quasi-experimental, not randomised
Comparison	Control group (n = 30); its instruction not described beyond the absence of the training	Comparison group taught by conventional audio-only methods
Intervention	Explicit instruction in connected-speech features (assimilation, linking, elision, vowel reduction)	Cooperative learning using YouTube clips versus conventional audio-only teaching
Duration	12 weeks; 24 sessions of about one hour within regular classes (p. 71)	Not stated in the consulted report
Outcome measure	Adapted LCPQ, pre and post, covering recognition, parsing, utilization	Pre-post listening comprehension scores; questionnaire; thematic analysis
Results	Pre-test between-group differences not significant (p = 0.387, 0.311, 0.347; p. 78); significant group x time interaction for all three problem categories (Wald chi-square 14.472-16.676, p < 0.001); treatment improved on all three, control unchanged	Significant rise in comprehension scores and motivation in the experimental group; no significant gain in the control group
Effect sizes and precision	d = 1.19 (recognition), 1.14 (parsing), 1.15 (utilisation) for the treatment group; d = 0.01-0.02 for the control group (pp. 80-83); calculation method not stated; no confidence intervals	Statistical test values reported; no confidence intervals
Perception versus proficiency	Perceptions of problems, not measured proficiency	Comprehension scores (performance-based) plus self-reported motivation
Limits of the evidence	Cluster-level random allocation of four classes with an undescribed procedure; selection not randomised; no blinding or fidelity check; cases without pre-test data eliminated and final analysed n not reported (p. 85); perception-based outcomes; same cohort as the baseline study	n = 10; no randomisation; confounding uncontrolled

Appendix I: Contextual Evidence Sources

Table 14. Contextual sources and their permitted use.

Source	Construct measured	Level	Permitted use
Pathan et al. (2016)	School ELT conditions	School	Background on preparation
Alghanimi & Ghazalla (2021)	Curriculum implementation, washback	School	Interpretive mechanism; not university evidence
Aldarasi (2020)	Foreign language anxiety (speaking-centred)	Higher education	Indirect affective background

Source	Construct measured	Level	Permitted use
Halali et al. (2026)	Vocabulary learning anxiety	Higher education	Adjacent construct; never estimates listening anxiety
Bin Naji et al. (2026)	Synthesis of 50 articles	Higher education	Comparison and interpretation only
Bahron & Abujela (2026)	Conceptual argument	Higher education	Argument and interpretation only