Journal of World Englishes and Educational Practices
ISSN: 2707-7586
DOI: 10.32996/jweep
Journal Homepage: www.al-kindipublisher.com/index.php/jweep

JWEEP
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A Systematic Self-Review of Specific-Skill Assessment Studies: Principles and Practices

Reima Al-Jarf
Full Professor of English and Translation Studies, Riyadh, Saudi Arabia
Corresponding Author: Reima Al-Jarf, **E-mail:** reima.al.jarf@gmail.com

| ABSTRACT

This study conducted a systematic review (SR) of the author's research program on specific language skill assessment published between 2004 and 2023. The SR includes 40 empirical and theoretical studies covering the assessment of listening, pronunciation, speaking, reading, writing, vocabulary, grammar, morphology, spelling, and research skills. The studies were organized into six thematic clusters: test construction principles; operationalizing and measuring process and product subskills; assessment in experimental studies; multi skill assessment; assessment instruments and technologies; and factors influencing skill assessment. Results of the SR showed that the studies provide a comprehensive, longitudinal, and practice based model for language skill assessment in EFL and L1 contexts. Across the studies, the author consistently demonstrated how process based instruction requires process based assessment, and how subskills must be operationalized, taught, and measured with precision. The findings also revealed that the author's test construction principles, such as alignment with instructional objectives, skill decomposition into process and product subskills, construct validity, reliability, and transparency, were applied consistently across all studies. Additionally, the SR showed that the author's experimental studies provided strong evidence that explicit instruction in subskills leads to significant gains in learners' performance, and that assessment instruments designed by the author were effective enough to detect such gains. Multi skill studies demonstrated the interconnectedness of language skills and the need for integrated assessment tasks. Studies on assessment technologies highlighted the author's early adoption of online tools, scoring iRubrics, and mobile apps to enhance assessment efficiency, fairness, and objectivity. Several contextual and learner related factors were found to influence assessment outcomes, including instructional consistency, teacher qualifications and expertise, test format clarity and logical sequencing, and learners' exposure to process and product subskill based training. The review also identified recurring pedagogical implications, emphasizing the need for systematic subskill instruction, alignment between teaching and testing, and the development of assessment literacy among teachers. Overall, the SR concluded that the author's research program constitutes a coherent, cumulative, and influential contribution to language skill assessment. It provides a replicable model for designing valid, reliable, comprehensive, discriminating and instructionally aligned assessment tools, and offers a rich foundation for future research on process based language assessment in similar educational contexts.

| KEYWORDS

Systematic review (SR), Al-Jarf research program, language skill assessment, assessment principles, assessment technologies, assessment tools, process subskill assessment, product subskill assessment, factors affecting assessment

| ARTICLE INFORMATION

ACCEPTED: 01 April 2026 **PUBLISHED:** 07 April 2026 **DOI:** 10.32996/jweep.2026.8.2.2

1. Introduction

Over the past century, language skill assessment has evolved from ancient, subjective oral examinations to highly scientific, communicative, and technology-driven methods. In the late 19th and Early 20th Century, the University of Cambridge Local Examinations Syndicate (UCLES) introduced the Certificate of Proficiency in English (CPE) in 1913, pioneering formal proficiency exams. The 1950s–1960s was the Psychometric-Structuralist Era, influenced by structural linguistics and behavioral psychology.

This era focused on "scientific" testing of discrete, individual language points (e.g., grammar, vocabulary). The Lado Test of Aural Comprehension (1946), developed by the English Language Institute at the University of Michigan, was the first American standardized tests. Large-scale tests like TOEFL (1960s) developed by the Educational Testing Service (ETS) revolutionized standardized language testing. The development of the Common European Framework of Reference for Languages (CEFR) has standardized proficiency measurement worldwide[1]. The 1970s–1980s witnessed the Communicative Approach Era. There was a shift from testing formal language features to assessing communicative competence, i.e., the ability to use language in real-world situations (Farhady, 2018). In the 2000s to the Present era, assessments moved from paper-based to internet-based and computer-adaptive formats (e.g., TOEFL iBT). This era emphasized formative feedback over just summative testing, learner-oriented assessment, interactional competence, and the use of technology for delivery. Modern trends emphasize assessment for learning, using data and feedback to help students improve rather than just ranking them. Current developments include automated scoring for writing and speaking and the use of AI to create more personalized, secure testing environments (Farhady, 2018).

Due to the long history of language assessment, an unlimited body of research has focused on different aspects of language learning and language skill development. However, systematic reviews (SRs) and meta-analyses (MSs) that address a variety of language skill assessments are recent and limited. The first group of SRs focused on broad or multi-skill evaluation rather than single-skill assessment such as forms of assessing language skills (Ardian et al., 2025); qualities of tests of English language skills in Indonesian schools (Djiwandono & Ginting, 2025); Glenn Fulcher's thirty-five years of contribution to language testing (Fulcher et al., 2022); self-assessment and language performance in language testing (Li & Zhang, 2021); quality in language assessment (Wind & Peterson, 2018); and psychometric properties of language assessments for children aged 4–12 years (Denman et al., 2017). These SRs examined general language assessment practices, psychometric properties, rating quality, and self-assessment. While these reviews provide valuable insights into overall assessment quality, they do not address the evaluation of individual language skills in L2 education.

SRs by El Kheir et al. (2023) and Bahi et al. (2024) focused on automatic pronunciation assessment systems and feedback mechanisms for Arabic learners. These reviews highlight advances in automated scoring but do not examine broader pedagogical or test-based approaches to assessing pronunciation as a single skill.

Further SRs focused on communication skills assessment in postgraduate medical training (Gillis et al., 2015); online crowdsourcing for assessing perceptual speech (Sescleifer et al., 2018); and tools for assessing speech-impaired children (Usha and Alex, 2023). Although relevant to oral communication, these SRs fall outside mainstream L2 language assessment and do not target single-skill evaluation in educational settings, as they focus on speech outcomes in clinical or medical contexts.

Some other SRs examined speaking assessment more directly. Saptiany and Prabowo (2024) reviewed speaking proficiency among English for specific purpose students. Marinho et al. (2022) explored public speaking assessment and self-assessment instruments. Hu et al. (2025) synthesized research on computer-based English speaking tests, and Taufiqi et al. (2025) reviewed e-assessment of speaking skills for advancing language evaluation in Indonesian language education. These studies contribute to understanding speaking assessment practices, yet they remain limited to specific contexts, technologies, or populations rather than offering a comprehensive synthesis of single-skill speaking assessment in L2 education.

Additionally, several studies reviewed expressive and receptive language skills, especially in early childhood or clinical populations such as universal strategies for improving expressive language skills in the primary classroom (Dobinson & Dockrell (2021); receptive and expressive English language assessments for young children (McIntyre et al. (2017); factors affecting receptive and expressive language in autistic children and siblings (Muès et al., 2024), and spoken language outcomes in children with ASD (Trembath et al., 2016). These studies are valuable but fall outside L2 assessment and do not address single-skill evaluation of normal students in educational contexts.

Reading assessment SRs received considerable attention from researchers such as digital reading comprehension assessment in English language education (Dang & Habók, 2026); dynamic assessment as a predictor of reading development (Dixon et al., 2023); randomized evaluation of reading skills (Honorato-Errázuriz & Ramírez-Montoya, 2021); reading tests (Ntonti et al., 2023); reading acquisition and analyses of the main international reading assessment tools (Sprenger-Charolles & Messaoud-Galusi, 2009), and the characteristics of dynamic assessment of word reading skills with implications for their validity (Wood et al., 2024). While informative, these reviews focus on specific tools, populations, or assessment types rather than providing a unified synthesis of single-skill reading assessment in L2 contexts.

[1] Language Assessment

Writing assessment SRs have primarily reviewed early writing tools and automated scoring systems (Buchanan et al. (2025) and automated writing evaluation systems as Grammarly, Pigai, and Criterion, with a perspective on future directions in the age of generative AI (Huawei & Aryadoust, 2023; Ding & Zou, 2024). These reviews highlight technological developments but do not address writing assessment as a single skill within broader L2 educational settings.

Only one review directly addressed grammar assessment. Chico (2026) conducted an MA on the effectiveness of grammar-integrated authentic assessment in enhancing language proficiency among high school students. Although relevant, this review focuses on instructional effectiveness rather than the systematic evaluation of grammar as an independent skill.

Regarding vocabulary assessment SRs, they have been reviewed in several SRs as vocabulary assessment in Brazilian children (Carbonieri & Lúcio, 2020), tools and teaching strategies for vocabulary assessment and instruction (Dujardin et al., 2021); the effectiveness of vocabulary size tests on students' English vocabulary and writing skills (Nitami & Santosa, 2024), and the dynamic assessment of vocabulary (Xia et al., 2024). These studies address vocabulary measurement but remain limited to specific populations or assessment approaches rather than offering a comprehensive synthesis of vocabulary assessment in L2 education.

Although the above SRs focused on specific skill assessment, such as speaking, writing, reading, vocabulary, or pronunciation, they tend to be narrow in scope, limited to particular tools, populations, or instructional approaches. None of the existing SRs provides a comprehensive, comparative synthesis of how specific language skills are assessed within mainstream L2 contexts. As a result, the field lacks an overarching understanding of what single-skill assessment looks like across various skills, how these assessments are designed, what constructs they target, and how they align with contemporary theories of language learning. This absence of an integrated, skill-by-skill synthesis represents a clear gap in the literature and underscores the need for the present SR. To fill this gap in the literature, the current study aims to conduct an SR of the author's studies on specific language skill assessment published between 2004-2023. The current SR covers 40 studies on language skill assessment principles, specific language skill assessment covering listening, speaking, pronunciation, reading, writing, vocabulary, grammar, morphology, spelling and research skills, skill assessment instruments and technologies and factors influencing skill assessment.

This study is significant because it is the first SR that provides a unique and comprehensive synthesis of a 20-year research program consisting of 40 classroom-based assessment studies conducted within a stable instructional context. It offers a holistic understanding of how process-based assessment principles were developed, refined, and applied across multiple cohorts, skills, and instructional cycles. It reveals patterns, consistencies, and cumulative contributions that were not visible when the studies were viewed in isolation. It provides a consolidated reference for educators and researchers interested in classroom-based assessment, particularly in contexts where instructors must balance heavy teaching loads with the demands of empirical research. It offers a model for long-term, instructor-led assessment research and underscores the value of sustained, context-embedded inquiry in improving language testing practices.

Furthermore, this SR is significant because it is part of a broader series of SR/MA projects by the author, that has so far cover the following: *teaching English for art education purposes to Ph.D. students (Al-Jarf, 2026a); themes, methods, and pedagogical insights in EFL reading instruction (Al-Jarf, 2026b); studies across diverse educational evaluation domains (Al-Jarf, 2026c); students' errors in English–Arabic and Arabic–English translation (Al-Jarf, 2026d); mobile apps for developing multiple language skills in EFL (Al-Jarf, 2026e); adult reading practices, interests, habits and challenges (Al-Jarf, 2026f); pronunciation instruction and practice in L2 (2005–2025) (Al-Jarf, 2026g); teaching reading in Arabic to grades 1–12: textbooks, skills, and learning outcomes (Al-Jarf, 2026h); Arabic–English transliteration of personal names and public signages (Al-Jarf, 2026i); children's language acquisition and development in Saudi Arabia (Al-Jarf, 2026j); classroom practices, writing enhancement and creativity among EFL struggling students (Al-Jarf, 2026k); collaborative learning and teaching in digital environments (Al-Jarf, 2026l); effectiveness of mind-mapping on multiple ENGLISH language skills in the Saudi context (Al-Jarf, 2026m); inadequate staffing and large class sizes in Saudi EFL and translation programs (Al-Jarf, 2026n); innovative word formation and pluralization processes in Arabic (Al-Jarf, 2026o); 2024–2025 studies on AI Arabic translation, linguistics and pedagogy (Al-Jarf, 2026p); three decades of ESP innovation across specialized and underexplored domains (Al-Jarf, 2026q).*

2. Methodology

2.1 Study Corpus

The study corpus comprises 40 empirical and descriptive studies published by the author between 2004 and 2023 in peer-reviewed journals, conference proceedings, book chapters, and research reports. Studies were included in this SR if they were authored or co-authored by *Reima Al-Jarf*, involved EFL learners, native Arabic-speaking students, or EFL teachers, and addressed the assessment of listening, pronunciation, speaking, reading, writing, grammar, vocabulary, morphology, spelling, research skills, or Arabic word-identification skills. Only studies published in English or Arabic and with full-text accessibility were eligible. The selected studies span the full trajectory of the author's research program and represent a wide range of assessment contexts,

instructional settings, and methodological approaches. For analytical coherence, the 40 studies were organized into six thematic clusters based on their primary assessment focus.

Cluster 1: Language skill test construction principles

Cluster 1 focuses on the principles of constructing tests that target specific language skills. The studies in this cluster examine how assessment instruments were designed, validated, and aligned with the process-based instructional approach. Each study demonstrates how test specifications (subskills and content covered), task formats, and scoring procedures were developed to measure learners' performance on clearly defined skill components. It includes the following studies:

1) *What teachers should know about reading tests (Al-Jarf, 2017b)*
2) *What teachers should know about vocabulary tests for EFL freshman students (Al-Jarf, 2015d)*
3) *Testing multiple vocabulary associations for effective long-term learning (Al-Jarf, 2023e)*
4) *Issues in assessing the speaking skill in EFL (Al-Jarf, 2015b)*
5) *How to prepare English language tests (Al-Jarf, 2009b)*

Cluster 2: Operationalizing and measuring process and product subskills

Studies in this cluster illustrate how complex language skills can be decomposed into measurable process subskills and assessed through task-based, purpose-specific, and diagnostically oriented instruments. Together, they demonstrate a systematic approach to defining subskills, designing process and product assessment tasks that capture both cognitive processes and observable products, and using empirical evidence to refine instructional and evaluative practices. This cluster highlights the author's sustained commitment to principled sub-skill measurement across diverse learner groups, course levels, and linguistic domains. Included studies are:

6) *Testing reading for specific purposes in an art education course for graduate students (Al-Jarf, 2021f)*
7) *Developing and testing reading skills through art texts (Al-Jarf, 2011b)*
8) *Assessing students' reading competencies: setting global standards (Al-Jarf, 2007b)*
9) *Testing reading for special purposes (Al-Jarf, 2007f)*
10) *Task-based instruction for EFL struggling college writers* (Al-Jarf, 2005a)
11) *Assessing graduate students' research skills in EFL (Al-Jarf, 2013a)*
12) *Testing research skills in EFL (Al-Jarf, 2007g)*
13) *freshman students' difficulties with English adjective-forming suffixes (Al-Jarf, 2019b).*
14) *Difficulties in learning English plural formation by EFL college students (Al-Jarf, 2022c).*
15) *Phonological and orthographic problems in EFL college spelling (Al-Jarf, 2008d).*
16) *Analysis of Arabic first, second and third grade students' errors in word identification (Al-Jarf, 1994)*

Cluster 3: Skill Assessment in Experimental Studies

Cluster 3 brings together a series of experimental studies that examine how technology-enhanced, collaborative, and blended instructional environments influence the assessment of core language skills. These studies investigate the effects of mobile learning, online platforms, and web-based instructional tools on learners' performance in reading, writing, vocabulary, grammar, and cultural awareness. By integrating assessment within controlled instructional interventions, the research demonstrates how process-sensitive and product-oriented measures can capture learners' development across diverse course types. Included studies are:

17) *Mobile technology and student autonomy in oral skill acquisition (Al-Jarf, 2012b).*
18) *the effect of web-based learning on struggling ESL college writers (Al-Jarf, 2004b);*
19) *Effects of online collaborative activities on second language acquisition (Al-Jarf, 2009a)*
20) *impact of blended learning on EFL college readers (Al-Jarf, 2007c)*
21) *Teaching vocabulary to EFL college students online (Al-Jarf, 2007e)*
22) *the effects of online grammar instruction on low proficiency EFL college students' achievement (Al-Jarf, 2005c);*
23) *impact of online instruction on EFL students' cultural awareness (Al-Jarf, 2006a);*
24) *collaborative mobile ebook reading for struggling EFL college readers (Al-Jarf, 2021)*
25) *collaborative mobile ebook reading by translation students (Al-Jarf, 2014)*
26) *Enhancing freshman students' performance with online reading and writing activities (Al-Jarf, 2013b);*
27) *integrating RCampus in college reading and writing for translation students (Al-Jarf, 2010b).*
28) *differential effects of online instruction on a variety of EFL courses (Al-Jarf, 2004a);*
29) *online ESL learning: effects on college levels & course types (Al-Jarf, 2007d);*

Cluster 4: Multi-skill assessment

Cluster 4 examines the interdependence of language skills and the ways in which performance in one domain influences outcomes in another. The studies in this cluster investigate how listening, decoding, background knowledge, and auditory processing interact with learners' spelling and interpreting performance. Together, they highlight the importance of assessing skills not in isolation but as interconnected components of a broader linguistic system. Included studies are:

30) *The effects of listening comprehension and decoding skills on spelling achievement of EFL freshman students Al-(Al-Jarf, 2005b);*
31) *effect of background knowledge on auditory comprehension in interpreting courses (Al-Jarf, 2018a).*

Cluster 5: Skill assessment instruments and technologies

Cluster 5 focuses on the development and use of assessment instruments and technology-enhanced tools designed to support language-skill evaluation. Studies in this cluster demonstrate how mobile applications, digital flashcards, and rubrics can enhance test preparation, standardize scoring, and improve the transparency of assessment practices. This cluster also includes diagnostic tools that target foundational skills such as word identification. Studies include:

32) *Standardized test preparation with mobile flashcard apps (Al-Jarf, 2021e)*
33) *Test preparation with mobile apps (Al-Jarf, 2014)*
34) *Creating and sharing iRubrics using RCampus (Al-Jarf (2010a)*
35) *Creating and sharing writing iRubrics (Al-Jarf, 2011a)*
36) *Creating and sharing vocabulary iRubrics (Al-Jarf, 2012a)*
37) *How EFL college instructors can create and use grammar iRubrics (Al-Jarf, 2020b)*
38) *An Arabic word identification diagnostic test for the first three grades (Al-Jarf, 1995)*

Cluster 6: Factors influencing skill assessment

Cluster 6 examines the contextual and human factors that shape the quality and effectiveness of skill assessment. The studies in this cluster highlight how instructor qualifications, pedagogical practices, and learner preferences influence assessment outcomes and the selection of appropriate evaluation methods. Studies include:

39) *Role of instructor qualifications, assessment and pedagogical practices in EFL students' grammar and writing proficiency (Al-Jarf, 2022i)*
40) *EFL female college students and instructors' preferred method of speaking assessment (Al-Jarf, 2021b).*

2.2 *Eligibility (Inclusion & Exclusion) Criteria*

Studies were excluded if they met any of the following criteria:

- **Duplicate studies that do not add new data or analysis.** Examples include: Grammar iRubrics for EFL teachers and students (Al-Jarf, 2011c); Reading for specific purposes in art education (Al-Jarf, 2021g); Word identification difficulties in early grades (Al-Jarf, 2018c); Background knowledge and auditory comprehension (Al-Jarf, 2018b); Online reading instruction for Arabic EFL learners (Al-Jarf, 2019c); Mobile technology and oral skill autonomy (Al-Jarf, 2011); Adjective-forming suffixes (Al-Jarf, 2008a); Effectiveness of internet-based EFL learning (Al-Jarf, 2008c); Plural acquisition by freshman EFL students (Al-Jarf, 2006b); Online learning and struggling ESL writers (Al-Jarf, 2002a); Phoneme–grapheme difficulties in EFL freshmen (Al-Jarf, 2019a); and Effects of e-learning on EFL instruction (Al-Jarf, 2006c).

- **Studies that assess advanced students' ability to identify the stylistic features of specialized texts,** even if they involve reading comprehension and specialized vocabulary meaning**.** These studies use a test to identify students' difficulties but they do not examine assessment design, validation, or construct definition. Examples are: can ESL students identify emphatic features of advertisements (Al-Jarf, 2025); problems of identifying lexical and syntactic features of legal documents by undergraduate EFL Students (Al-Jarf, 2023c; Al-Jarf, 2021c;); processing of advertisements by EFL Arab college students (Al-Jarf, 2007); translation students' difficulties with English neologisms (Al-Jarf, 2010c).

- **Studies on translation assessment.** These studies evaluate translation performance rather than a single language skill. Examples include: grade inflation in language and translation courses (Al-Jarf, 2022h); critical analysis of translation tests (Al-Jarf, 2021a); analytical and measurement issues in translation tests (Al-Jarf, 2001; 2002a; 2002b; 2003); studies on translation difficulties involving cultural or linguistic expressions (Al-Jarf, 2016; 2017a; 2019b; 2022; 2022f; 2023b; 2023d;

2023f); online exams in language, linguistics and translation courses (Al-Jarf, 2022h); expressions of impossibility in Arabic and English (Al-Jarf, 2024).

- **Studies on general educational evaluation domains**: These studies do not assess a specific language skill. Examples include: thesis evaluation, grade inflation, peer-reviewing, instructor performance, textbook evaluation, program evaluation, material coverage, admission tests, and staffing benchmarks (Al-Jarf, 1989; 1991; 1998; 2007a; 2008b; 2008e; 2015a; 2020a; 2021d; 2022e; 2022g; 2023b).

- **Studies on journal article assessment and peer-review challenges.** Examples include: challenges faced by Arab peer-reviewers (Al-Jarf, 2023a); challenges faced by peer-reviewers in academic institutions (Al-Jarf, 2019); Evaluation Checklists in isolation without a paper: *textbook evaluation checklist (Al-Jarf, 2015c).*

- **Evaluation checklists presented in isolation without an accompanying study.** Example: textbook evaluation checklist (Al-Jarf, 2015c).

- **Skill assessment in non-academic domains:** *deviant Arabic transliterations of foreign shop names in Saudi Arabia and decoding problems among shoppers (Al-Jarf, 2022b).*

2.3 *Corpus Characteristics*

The final corpus consisted of 40 studies authored by Reima Al-Jarf between 2004 and 2023. Because the dataset represents a closed, author-bounded research program focused on language-skill assessment across different instructional contexts, learner populations, and assessment purposes, it is both comprehensive and internally coherent, reflecting the author's sustained scholarly trajectory in assessing speaking, reading, writing, vocabulary, grammar, and research skills. Although diverse in their specific topics, the studies share a common orientation toward testing language skills, examining assessment principles, and exploring tools, technologies, and pedagogical conditions that shape assessment practices. The corpus spans a range of publication sources, including conference proceedings, journal articles, and academic presentations, and covers multiple skill domains such as listening, speaking, reading, vocabulary, grammar, research, and integrated-skill assessment. As a temporally bounded corpus, it provides a stable dataset that allows for tracing conceptual development, methodological patterns, and thematic continuity across the author's work. This structure also supports a focused synthesis aligned with the aims of the present review.

2.4 *Information Sources*

The information sources for this SR were limited to platforms that index the author's complete scholarly output. No external database search was required, as the aim was not to identify all global studies on educational evaluation, but rather to synthesize all studies related to language skill assessment within a single, self-contained research program. All records were retrieved from publicly accessible academic platforms in which the author's publications are archived. These sources include Google Scholar, ERIC, ResearchGate, Semantic Scholar, Academia.edu, SSRN, EBSCO, ProQuest, and institutional repositories. Collectively, these platforms provide full coverage of the author's publications across journals, conference proceedings, reports and digital repositories. All included and excluded studies were verified manually to ensure accuracy, remove duplicates, and confirm alignment with the eligibility criteria described in Section 2.2.

2.5 *Data Extraction and Synthesis*

Data from all eligible studies were systematically extracted using a structured extraction sheet developed for this review. For each study, information was recorded on publication year, research purpose, target skill(s), assessment principles, tools and technologies used, methodological features, participant characteristics, instructional context, and key findings related to language skill assessment. All extracted information was cross-checked against the full text of each study to ensure accuracy and completeness. Following extraction, the studies were synthesized using a thematic narrative approach appropriate for a heterogeneous corpus spanning descriptive, diagnostic, and experimental designs. Recurring patterns, converging findings, and conceptual overlaps were identified and organized into seven thematic clusters representing: (1) general principles of language skill assessment; operationalization and measurement of process and product subskills; assessment in experimental studies; multi-skill assessment; assessment instruments and technologies, and factors influencing assessment practices. This synthesis approach enabled the integration of insights across diverse studies while preserving the distinct methodological and conceptual contributions of each publication. Because the corpus represents a single author's long-term research program, the methodological framing and

analytical categories were highly consistent across studies, which minimized coding discrepancies and facilitated a unified synthesis of findings spanning nearly two decades of scholarly work.

2.6 *PRISMA Flow Description*

Because this SR is based on a closed, predefined corpus of 40 studies published by the same author between 2004 and 2023, the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow reflects a streamlined and highly controlled identification and screening process. All publications produced within this period were retrieved from the academic platforms listed in Section 2.4 and manually screened for relevance. Each record was evaluated against the eligibility criteria, and studies were excluded if they were duplicates or if they focused on areas outside the scope of this review—such as translation studies, general educational evaluation, program evaluation, textbook analysis, peer reviewing, grade inflation, admission testing, Arabic reading programs, or isolated checklist development without an accompanying research investigation. Following full-text screening, only studies that directly addressed language skill assessment principles, the operationalization and measurement of specific language skills, assessment in experimental studies, multi-skill assessment, assessment instruments and technologies, and factors influencing assessment practices were retained. The final set of eligible studies was then organized into 6 thematic clusters for synthesis. Accordingly, the PRISMA flow documents the progression from the initial identification of all publications within the author-bounded corpus, through systematic screening and eligibility assessment, to the final inclusion of studies that substantively contribute to the core assessment domains of this review.

3. Results

3.1 Overview

This SR synthesizes findings from 40 studies on language skill assessment conducted over nearly two decades. The analysis is organized into 6 thematic clusters, allowing the results to highlight both the distinct contributions of individual studies and the cross-cluster patterns that characterize the author's broader research program. Across the corpus, the studies consistently examine general principles of language skill assessment, the operationalization and measurement of process and product subskills, assessment in experimental interventions, multi-skill assessment, assessment tools and technologies, and the factors that affect assessment practices. This overview presents the major trends emerging from the corpus and illustrates how each thematic cluster contributes to a deeper understanding of language skill assessment. Collectively, the studies document assessment practices, methodological approaches, and instructional conditions across a wide range of skill domains, including listening, speaking, reading, writing, vocabulary, grammar, research skills, and word-identification processes.

3.2 Study Characteristics

The corpus consisted of 40 unique studies distributed over 7 thematic clusters. Findings of each cluster are presented below.

Cluster 1: General skill assessment principles

How to prepare English language tests (Al-Jarf, 2009b)

The article presents comprehensive guidelines for designing and evaluating language skill tests within translator and interpreter training programs. It emphasizes that English courses in levels 1–4 must prepare students for advanced translation for the translation and interpreting courses that require high proficiency levels. The article outlines specific subskill for: (i) *Listening*: main ideas, details, phoneme and allophone recognition, stress, intonation, reductions, note-taking, summarizing. (ii) *Speaking*: idea generation, details, organization, fluent speech with correct pronunciation, grammar, and vocabulary. (iii) *Reading*: main ideas, details, text structure, vocabulary inference, pronoun reference, idioms, derivatives, summarizing, outlining. (iv) *Writing*: coherent paragraphs and essays with correct grammar, cohesion, and organizational markers. (v) *Vocabulary*: pronunciation, spelling patterns, stress, part of speech, countability, derivatives, idioms, collocations, and English and Arabic meanings. The test instructions should be brief and explicit. Questions should be mainly productive and require students to think critically, infer, synthesize information, and perform tasks at the word, sentence, paragraph and discourse levels. It gives guidelines for test length, timing, and duration, and specifies formatting specifications for the test paper, margins, font size, spacing, headers, and stresses that tests should remain plain and free of decorative elements.

Issues in assessing the speaking skill in EFL (Al-Jarf, 2015b)

The article discusses key considerations in administering and scoring speaking tests in both language-lab and face-to-face settings. Test instructions must clearly specify the required details, task type, and linguistic features (sentence structure, pronunciation, stress, intonation, and fluency). Test topics should be comparable to, but not identical with, classroom practice topics. In lab testing, students receive printed questions and parallel versions. Students write their names on the question paper, practice using their MP3 players at home, test the device at the beginning of the session, read the questions, think about them, and then record their responses in any order while stating the question number. Students are not allowed to write answers and read them aloud, nor to re-listen, erase, or re-record their responses. During scoring, the teacher listens to the recordings and takes notes directly on the

student's question paper, noting strengths and weaknesses. In face-to-face exams, students are tested individually. A student draws printed questions from a basket; each student completes several speaking tasks within 10–15 minutes. The speaking scoring rubric consists of three skills (idea generation or content, grammar and vocabulary, and pronunciation and fluency), and three performance levels (excellent, average, poor), with descriptors and mark ranges specified for each level.

Testing multiple vocabulary associations for effective long-term learning (Al-Jarf, 2023e)

This article emphasizes that vocabulary tests should measure students' ability to think, apply, infer, connect, and synthesize information about words and phrases rather than relying on simple recall. It outlines a wide range of vocabulary subskills that should be incorporated into assessment, including pronunciation, spelling, part of speech identification, morphological structure, semantic aspects, register, and usage. The article details specific tasks such as recognizing silent letters, hidden consonants, double letters, vowel–sound inconsistencies, syllabication, and stress; spelling words correctly; defining words in English and Arabic; identifying opposites, synonyms, and related words; grouping words by meaning; determining part of speech; supplying appropriate prepositions; applying capitalization rules; identifying singular and plural forms; selecting verbs that collocate with particular nouns; distinguishing count and non-count nouns; and recognizing abstract versus concrete nouns and American versus British usage. It also highlights the importance of identifying derivatives, prefixation, suffixation, forming compounds, recognizing and using words, idioms, collocations, prepositional phrases, and phrasal verbs in correct sentences, distinguishing idiomatic expressions from non-idiomatic ones, ensuring that vocabulary assessment captures the full range of linguistic, semantic, and functional associations necessary for long-term learning.

What teachers should know about reading tests (Al-Jarf, 2017b)

This study identifies the essential principles that teachers need to understand when designing and evaluating reading tests: (i) reading product and reading process skills, text components, reading skills, reading comprehension levels; (ii) types of tests: aptitude, diagnostic, achievement; formative vs summative assessment; speed vs bower test; (iii) characteristic of a good test (item difficulty level and discrimination power, test validity and reliability); (iv) identifying the content to be covered by the test and the reading skills to be tested (main ides and supporting details explicitly or implicitly stated, recognizing text structure, inferring meanings of difficult words from context; connecting pronouns with their antecedents, skimming, scanning… etc.); selecting the reading passage length, difficulty level, topic; test duration; number of tests per semester; preparing the table of specifications (skills, content, total test items, marks allocated); questions format (short answer, multiple choice, true-false); arrangement of test items; preparing the test draft; test instructions (should be clear, brief, specific, simple language); test paper format (margins, font, line spacing, pagination, instructions, teacher's name, date, interm number, and course); preparing an answer key; and scoring the test.

What teachers should know about vocabulary tests for EFL freshman students (Al-Jarf, 2015d)

This article provides detailed guidelines for planning and designing vocabulary tests that include: skills to be covered (Pronunciation, spelling, identifying the part of speech, morphological structure, semantic aspects, register); outlining the course content; preparing a table of specifications showing the skills, content topics and number of questions allocated to each; writing brief and clear instructions. Questions should cover all skills, tasks and exercises covered in the classroom and textbook. The test tasks should cover word, sentence, paragraph & discourse levels; test student's ability to think, apply, infer, connect, and synthesize information, not mere recall; more production than recall questions, should be reliable and valid, have adequate discriminating power; should be a power and a speed test, and have an average difficulty level. The article also shows the test length (total number of words, how many pages, how many words per question), when to give the tests and test duration; the test paper format (margins, font size, instructions, line spacing, pages, numbering pages, and items, test and student information, simplicity, no decorative boxes, circles, flowers); an answer key and how tests are scored.

Cluster 2: Operationalizing and measuring process and product subskills
Assessing students' reading competencies: setting global standards (Al-Jarf, 2007b)

This study proposed a model for testing reading for any purpose, any level and any text type. It serves as a general framework for EFL instructors. The model is based on the reading product and process theories. A reading test consists of an unseen text of a length, difficulty level and type comparable to those practiced in class. The reading text should be followed by items that measure the following product and process skills: literal and inferential comprehension, evaluation, and appreciation, using phonics clues, using word structure clues, using semantic or contextual clues, using syntactic clues to determine the meaning of unknown words, recognizing text structure, making inferences, and recognizing anaphoric relationships. Recognizing types of cohesion, the test items should cover all 12-skill areas. The subskills can be chosen in accordance with the students' proficiency level. Questions should not involve direct copying of answers from the text.

Testing reading for specific purposes in an art education course for graduate students (Al-Jarf, 2021f)

Based on the results of a needs assessment questionnaire and an English proficiency test, an ESP course was especially designed to meet art education students' academic and professional needs. At the end of the semester, the students were posttested. The posttest required the students to locate main ideas and supporting details and to figure out the meanings of key art terms from context. It also required them to identify the part of speech of art terms and detach suffixes. The test also included items that asked students to provide the overall meaning of short paragraphs and single sentences in Arabic. The article presents a detailed description of the content sampled, the reading skills assessed, examples of test items, and the statistical analyses conducted on the ESP reading test scores, illustrating how the assessment was aligned with the students' academic and professional needs.

Developing and testing reading skills through art texts (Al-Jarf, 2011b)

Based on a needs assessment questionnaire and an English Proficiency Test result, a reading course was especially designed to meet art education students' academic, professional and communication needs. At the end of the semester, the students were post-tested. A detailed description of the reading course: skills developed, art materials selected, reading test developed, reading skills tested, and results of the training are reported.

Testing reading for special purposes (Al-Jarf, 2007f)

The ESP posttest was designed to measure graduate students' ability to read and comprehend English art-related texts, assess their vocabulary knowledge, and evaluate their ability to translate art terminology and overall text meaning into Arabic. The reading component targeted recognition of macro- and micro-structures, identification of main ideas and supporting details, locating specific information, skimming online materials such as art websites and Amazon book citations, and identifying key concepts and terms. Vocabulary skills included inferring meanings of art terms from context, identifying parts of speech, and recognizing prefixes and suffixes. Translation skills assessed students' ability to translate individual words as well as the overall meaning of an art text. The test content closely matched classroom materials in theme, difficulty, and length. It included a long biographical text about Picasso, several short art texts, an art gallery text, book citations, and isolated vocabulary items. A detailed table of specifications showed the distribution of items across reading, vocabulary, and translation skills, ensuring alignment with instructional objectives. Results indicated a significant improvement in students' performance, demonstrating the effectiveness of the ESP course. There were strong positive correlations between vocabulary and translation scores, and between reading and translation scores. The correlation between reading and vocabulary was not significant. The test showed strong discrimination power between students who mastered the targeted skills and those who did not.

***Task-based instruction for EFL struggling college writers* (Al-Jarf, 2005a)**

Pretest writing results of 65 EFL freshman students showed that the students could not put two words together. The posttest results showed a great improvement in writing ability. The students could write fluently and communicate easily. Spelling, punctuation and capitalization errors significantly decreased. Improvement was noted in essay length, neatness, mechanical correctness and style. Improvement was due to student and efficient task management factors such as accepting comments on their essays, practicing weekly writing tasks under supervision, individual help, extension activities, putting all tasks together in writing a one-paragraph essay, encouraging writing without worrying mistakes, communicative feedback focusing on meaning and only errors related to tasks under study, giving feedback on the presence and location of errors, self-editing and peer-editing, giving extra credit, giving quizzes every other week, returning graded quizzes with comments on strengths and weaknesses, with words of encouragement and discussing answers in class.

Assessing graduate students' research skills in EFL (Al-Jarf, 2013a) and Testing research skills in EFL (Al-Jarf, 2007g);

A research training module was designed to help graduate students majoring in art education to locate, read and comprehend art abstracts and full-text articles for their assignments and theses. At the end of the training module, the students were post-tested. They were given individual research projects, for which they had to select the search terms, define the search strategy, go online, log into the electronic database, conduct simple and advanced searches, and print and save the actual records obtained. The posttest required the students to locate art dissertation abstracts and journal articles etc. The students were asked to skim through sample art abstracts and journal articles and locate the aim of the study, the type of instrument used in collecting the data, the subjects, data collection procedures, statistical analysis procedures and results and give a summary translation in Arabic. A detailed description of test content and the research skills measured are given.

Freshman students' difficulties with English adjective-forming suffixes (Al-Jarf, 2019b)

EFL freshman students at the College of Languages and Translation (COLT) received direct instruction in adjective-forming suffixes, and then they took an immediate and a delayed test. Analysis of 547 errors collected from responses to test showed that 36% of

the responses were left blank or the subjects duplicated the stimulus word. In 32% they mismatched the word and suffix, in 36%, they made spelling mistakes; in 15% they spelled words phonetically, and in 15% they added a noun- or an adverb-forming suffix. Significant differences were found in the number of errors made on both tests which correlated with the students' vocabulary knowledge. A hierarchy of difficulty in attaching adjective-forming suffixes, faulty strategies used in adjective morphology and possible causes of students' difficulty in adjective suffix acquisition are reported.

Difficulties in learning English plural formation by EFL college students (Al-Jarf, 2022c)

3099 plural formation-errors were collected from an immediate test a week after instruction and a delayed test at the end of the semester. No significant differences were found in the amount and types of errors made by the students in the immediate and delayed tests. students tended to regularize English plural formation and overgeneralize regular English plural morphemes (63.28%), i.e., they deleted the regular plural suffix from nouns ending in an -s or -es (35.37%) or tended to add the regular plural suffix to words that do not have it (27.91%). They confused singular and plural endings of Latin words (15.07%). they thought the singular and plural forms of a word were the same (7%). The most difficult plurals to master were those of words that end with an -s or -es but have no singular form (28.85%); words with Latin plurals (21.85%); non-count nouns with no plural form (21.4%), and words that have a plural, but they thought they had no plural form (8.55%). Interference among the English plural morphemes themselves and confusing plural formation rules caused most errors. No interference from Arabic pluralization was found.

and orthographic problems in EFL college spelling (Al-Jarf, 2008d)

36 Saudi EFL freshmen students, at the COLT, took a listening-spelling test in which they filled out 100 blanks in a dialogue. Results indicated that 63% of the spelling errors were phonemic and 37% were graphemic. It was also found that the subjects had more problems with whole words than with graphemes and phonemes. Some of the phonemic problems that the subjects had were inability to hear and discriminate all or most of the phonemes in a word, inability to discriminate vowel phonemes and hear the final syllable or suffix. They mostly had graphemic problems with vowel digraphs, double consonants, silent vowels and consonants, and homophones. A simplification process seems to affect students' spelling errors. A detailed account of EFL students' phonemic and graphemic errors in spelling is given.

Analysis of Arabic first, second and third grade students' errors in word identification (Al-Jarf, 1994)

A word-identification diagnostic test was used to assess Grade 1, 2 and 3 students' weaknesses in auditory and visual discrimination, letter recognition, sight word recognition, word recognition in context, sound-symbol association, and structural analysis subskills. Students' responses revealed that the easiest subskills for all grades were visual and auditory discrimination, sight word recognition or word recognition in context; and to a moderate degree: letter recognition; and the most difficult subskills were sound-symbol association and structural analysis respectively. The three grades were significantly different in their mean error in word, identification in general and in the different subskills. The Grade 1 mean error was greater than Grade 2 and 3 mean error in auditory and visual discrimination, letter recognition and word recognition in context, but there were no significant differences between the second and third grade mean error. The Grade 1 mean error was greater than the Grade 2 which was in turn greater than the Grade 3 errors in sight word recognition, sound-symbol association and structural analysis and the total number of errors. Students' mastery of word identification in general and of each sub-skill gets better as they proceed from one grade to the next. The correlation between the total error score and errors in each sub-skill was positive, significant, and greater than the correlation between the errors in the subskills.

Cluster 3: Assessment in Experimental Studies
Mobile technology and student autonomy in oral skill acquisition (Al-Jarf, 2012b).

Pre-test scores showed no significant differences in oral proficiency level between the experimental and control groups. Both groups were exposed to the same in-class instruction that depended on the textbook. The experimental group used a self-study MP3 English listening and speaking program. They mostly covered 90 lessons and listened to 900 short audio files of Basic English structures and commonly used expressions out of class. The MP3 lessons consisted of short sentences which the students could read, listen to and mimic as many times as they needed. On average, the students practiced 3.5 hours a week. The posttest showed significant differences between the experimental and control groups as a result of using the MP3 self-study lessons. The experimental group made higher gains in listening and speaking abilities manifested in listening comprehension, oral expression, fluency, pronunciation correctness and vocabulary knowledge. There were positive correlations between practice time and the number of lessons covered and between listening and speaking posttest scores. Students reported positive attitudes towards the MP3 self-study listening and speaking lessons and reported several benefits.

The effect of web-based learning on struggling ESL college writers (Al-Jarf, 2004b); impact of blended learning on EFL college readers (Al-Jarf, 2007c); Teaching vocabulary to EFL college students online (Al-Jarf, 2007e); the effects of online grammar instruction on low proficiency EFL college students' achievement (Al-Jarf, 2005c); impact of online instruction on EFL students' cultural awareness (Al-Jarf, 2006a)

In each of these studies, pretest results showed no significant differences between the experimental and control groups in achievement and skill level. Then, the experimental group received a combination of in-class instruction that depended on the textbook and online instruction. Depending on the course, the students used platforms like Nicenet or Blackboard to summarize texts, identify details, analyze vocabulary in context, classify information, break cultural terms into their component parts, complete online exercises and quizzes, respond to discussion threads, locate information from external websites, post short paragraphs, stories, or poems on the discussion board and check the links posted by the instructor, post responses, respond to each other, and correct their own errors. Regardless of the skill taught, posttest results consistently showed that students who completed online activities made significantly higher gains than those who relied on traditional instruction only.

Collaborative mobile ebook reading for struggling EFL college readers (Al-Jarf, 2021); collaborative mobile ebook reading by translation students (Al-Jarf, 2014); effects of online collaborative activities on second language acquisition (Al-Jarf, 2009a)

In both studies, pre-test results showed no significant differences between the experimental and control groups in reading ability. In addition to in-class instruction that depended on the textbook, the experimental group engaged in extensive collaborative e-book reading activities. Experimental students worked in small groups, and were assigned a free e-book in one study and weekly topics (themes) – in the other study - to investigate such as types of natural disasters. Each group selected a sub-topic, searched and synthesized information about the sub-topic (such as tsunami in Indonesia, Hurricane Katrina, earthquakes in Iran), posted a paragraph about the subtopic and asked questions about the main idea of their paragraph, details, guessing meanings of difficult words from context, connecting pronouns with their referents and so on. Each group read a chapter (2-3 pages), posted an outline or summary, questions, answers, feedback, comments, study skills and self-improvement tips and websites on the Discussion Forum of Nicenet or the Blackboard and held discussions via Elluminate web conferences. Post-test results in both studies showed that the experimental groups outperformed the control groups. They made significantly higher gains in reading comprehension and overall language skills due to student-centered activities, real-life concrete topics, topics of interest for the students, students encouraged to express themselves, active participation and practice, clear instructions, a secure environment for making mistakes, and instructor and peer support.

***Enhancing freshman students' performance with online reading and writing activities (Al-Jarf, 2013b); integrating RCampus in college reading and writing for translation students* (Al-Jarf, 2010b).**

Pretest results in both studies showed no significant difference between the experimental and control groups in their reading and writing skills. In addition to in-class textbook-based instruction, the experimental group received online instruction using RCampus. Each week, discussion threads that required the students to search for information, read extra material and respond to questions in writing were posted. The students were free to post their own book summaries, discussion threads and comment on each other's posts. Comparisons of the pre- and posttest mean scores revealed that freshman students' reading and writing skills significantly improved as a result of using a combination of online instruction via RCampus and traditional in-class reading and writing instruction using the textbook. Students' responses showed improved comprehension and production of main ideas, and supporting details that are explicitly or implicitly stated in the text; guessing meanings of difficult words from context, explaining meanings in their own words; connecting pronouns with their antecedents and writing a summary and making an outline of the main ideas and most important supporting details in the text.

Differential effects of online instruction on a variety of EFL courses (Al-Jarf, 2004a)

The author taught 4 types of EFL courses to undergraduate students online: Grammar, writing, culture and study skills using Blackboard and Nicenet from home. Significant differences were found between pre- and post-test scores in writing, grammar and culture but not in study skills. The achievement level was higher among active participants who posted threads and shared in the discussion than among passive participants who were just browsers and did not write anything, and between members of the latter group and those who were not registered in the online courses at all. The effect of online instruction on students' attitudes is also reported.

Online ESL learning: effects on college levels & course types (Al-Jarf, 2007d)

1601 female students participated in the study. They were in the first, fourth, fifth and sixth semesters in college and some were Ph.D. students. The subjects were enrolled in 7 undergraduate ESL courses (reading, writing, grammar, vocabulary, culture, study skills and translation) and an English-for-specific-purposes (ESP) graduate course. In each course, students were divided into an

experimental and a control group. Control groups received textbook-based instruction, whereas experimental groups received a combination of traditional and online instruction using Blackboard or Nicenet from home. In each online course, discussion threads and internet resources related to the topics, skills and structures taught in class were posted. In each course, the students were pre and post-tested. Significant differences were found between pre- and post-test scores in reading, writing, grammar, vocabulary, culture and ESP courses but not in translation and study skills courses. The achievement level was higher among active participants who posted threads and shared in the discussion than among students who were just browsers and did not write anything, and between members of the latter group and those students who were not registered in the online courses at all. Online instruction had a positive effect on experimental students' achievement regardless of the type of LMS used. Online instruction had a positive effect on experimental students' achievement in all college levels (graduate and undergraduate; lower and upper undergraduate classes) and had a positive effect on experimental students' attitudes towards the reading, writing, grammar, vocabulary, cultures and ESP courses and online instruction. The study concluded that in learning environments where technology is unavailable to graduate and undergraduate EFL students and instructors, use of online courses from home and even as a supplement to in-class techniques helps motivate and enhance EFL graduate and undergraduate students' learning and development of their reading, writing and ESP skills, and grammatical, vocabulary and cultural knowledge in English as a second language.

Cluster 4: Multi-skill assessment

The effects of listening comprehension and decoding skills on spelling achievement of EFL freshman students (Al-Jarf, 2005b)

Thirty-six EFL freshman students at COLT were given a dictation, a listening comprehension test and a decoding test. Results showed that the typical EFL freshman student misspelled 41.5% of the words on the dictation, gave 49.5% correct responses on the listening comprehension test, and 52% correct responses on the decoding test. The subjects' spelling, listening and decoding achievement was low, which implied that the subjects were having spelling, listening comprehension and decoding difficulties. There were strong correlations between spelling ability, listening comprehension and decoding skills. This means that good spelling ability in EFL is related to good listening comprehension and good decoding skills. The better the listening comprehension and decoding abilities, the fewer the spelling errors. When listening comprehension and decoding skills are poor, spelling ability is also poor. Recommendations for spelling, listening and decoding instruction are given.

Effect of background knowledge on auditory comprehension in interpreting courses (Al-Jarf, 2018a).

Results of an interpreting pre-test showed that students majoring in an interpreting course at COLT have problems with media reports. They have difficulty discriminating phonemes and comprehending the meaning of unfamiliar foreign proper nouns such as place names, names of politicians, organizations, chemicals or diseases that they encounter in oral media reports. One month after the beginning of the semester, the subjects took an interpreting test which consisted of 5 Arabic and 5 English dialogues consisting of media reports on education, IT, politics, medicine and business topics. The dialogues were tape-recorded. The test was given in the language lab. Each student recorded her interpretation on a tape recorder. Analysis of 560 errors in interpreting foreign proper nouns such as place name, names of politicians, organizations, news agencies in the dialogs revealed a significant correlation between the students' overall interpreting accuracy score and her vocabulary errors score, which means that good student interpreters rendered highly accurate interpretations of the dialogues and made fewer vocabulary meaning errors, whereas poor student interpreters produced poor, incomprehensible and incoherent interpretations of the dialogues and many vocabulary meaning errors. student interpreters had problems comprehending media reports and interpreting their content from English into Arabic and Arabic into English. The students had difficulty in discriminating phonemes of unfamiliar foreign proper nouns such as place name, names of politicians, organizations, and news agencies that they encounter in media reports. They had difficulty comprehending the meaning of unfamiliar chemicals, diseases, names of organizations, measurement units, acronyms referring to international organizations, political posts and providing the correct English or Arabic equivalent.

Cluster 5: skill assessment Tools and technologies

Standardized test preparation with mobile flashcard apps (Al Jarf, 2021e) and Test preparation with mobile apps (Al Jarf, 2014d)

Across these two studies, mobile flashcard apps (FCAs) were examined as supplementary tools for preparing EFL college students for standardized tests such as the IELTS, TOEFL, TOEIC, GRE, and SAT. These apps, which are freely available on Google Play, the Apple App Store, and other mobile platforms, contain thousands of essential and specialized vocabulary items across diverse academic fields. They offer multiple learning modes (Study, Slide Show, Matching, Memorize, Quiz, and Play), customizable features such as starred words, and flexible browsing options. Both studies provide examples of FCAs, guidelines for locating and selecting appropriate apps, and a structured instructional model consisting of pre-task, task, and post-task phases. Instructors act as facilitators by helping students choose suitable apps, providing guiding questions, and monitoring progress. Together, the studies

demonstrate that mobile flashcards are accessible, efficient tools that support faster vocabulary learning and enhance students' readiness for standardized tests.

Creating and sharing iRubrics using RCampus (Al-Jarf (2010a); creating and sharing writing iRubrics (Al-Jarf, 2011a); creating and sharing vocabulary iRubrics (Al-Jarf, 2012a); how EFL college instructors can create and use grammar iRubrics (Al-Jarf, 2020b); empowering EFL teachers and students with grammar iRubrics (Al-Jarf, 2011c)

This series of studies presents iRubric as a comprehensive digital tool for designing analytic scoring rubrics for writing, vocabulary, and grammar courses by defining the skills, subskills, performance levels, and mark allocations for each course. Instructors can attach rubrics to coursework so that the students clearly understand expectations and can use them for self-assessment. The studies demonstrate how teachers can build, edit, apply, and share rubrics through the RCampus LMS, with scores automatically calculated and posted to the gradebook. They highlight the advantages of iRubric, including clearer performance criteria, alignment with learning outcomes, time-saving scoring procedures, secure access to scored rubrics, and opportunities for collaborative assessment through the RCampus rubric gallery. Collectively, these articles consider iRubric as an effective tool for improving the reliability and validity of EFL skill assessment.

An Arabic word identification diagnostic test for the first three grades (Al-Jarf, 1995)

This study developed a diagnostic test to identify symbol and word recognition difficulties in Grades 1–3. The test included seven components covering prereading readiness, decoding, and recognition skills (auditory and visual discrimination), while the remaining five assess letter–sound correspondence, word recognition in isolation, word recognition in context, letters-sound associations, and structural analysis). Test content was based on a detailed analysis of the Arabic writing system, spelling rules, and curriculum materials. A pilot with 633 students demonstrated high reliability and strong content validity. The study provides a comprehensive diagnostic tool for early reading assessment. The test package consists of 3 books: text construction, a teacher's guide and the test booklet.

Cluster 6: Factors influencing skill assessment

Role of instructor qualifications, assessment and pedagogical practices in EFL students' grammar and writing proficiency (Al-Jarf, 2022i)

The study examines how instructional and assessment practices influence the grammar and writing performance of EFL freshman students enrolled simultaneously in both courses. Three groups were compared: one group was taught grammar and writing by the same instructor, while the other two groups were taught by different instructors using the same textbook but different instructional and assessment techniques. Post-test results showed strong correlations between grammar and writing scores, with the highest achievement recorded among students taught both courses by the same instructor, suggesting a reciprocal relationship between grammatical competence and writing development. The study argues that when one instructor teaches both courses, she can make clearer and more effective connections between the content and skills required in each. Instructor qualifications, pedagogical systems, professional experience, integration of online instruction, types of error correction and instant feedback, and the use of formative assessment techniques were all found to be significantly more effective than instruction that relied solely on the textbook. These variables played a crucial role in improving the grammatical knowledge and writing quality of low-ability EFL students and resulted in substantial gains in their grammar and writing post-test scores.

EFL female college students and instructors' preferred method of speaking assessment (Al-Jarf, 2021b);

The study examines students' and instructors' preferences for two speaking assessment methods used at COLT: face-to-face testing and language-lab testing, and the reasons behind these preferences. Face-to-face assessment is the most common method, where students are tested individually for 5-10 minutes. They draw a topic from a basket, and engage in a short conversation or interview with the instructor, with testing sessions running continuously from morning until early afternoon. In the lab-testing method, all students take the test simultaneously; they read printed questions, take brief notes, and record their responses on MP3 players within an hour. The findings showed that most students prefer lab testing because the questions are more comprehensive, the conditions are uniform, they are less anxious, and if they miss a question, they do not lose a lot of marks. Unlike students, instructors prefer face-to-face assessment because questions are easier, cover only part of the material, and often result in all students passing the course, while students feel more anxious. Comparisons of test scores indicate that lab speaking tests are more reliable, more valid, and have stronger discriminating power between students who have mastered and those who have not mastered the speaking skills. The study concludes with recommendations for improving speaking assessment practices.

Although the studies in this cluster explicitly examine contextual or pedagogical factors influencing skill assessment, it is important to note that *all* studies in the corpus, across all the clusters, identify factors that contribute to students' improvement. In several experimental and task-based studies (e.g., Al-Jarf, 2005a), such factors are embedded within the instructional intervention rather than being the primary focus of the research.

3.3 Post-hoc Note

It is noteworthy to say that the Interactions I & I and Mosaic I & II textbook series assigned for teaching listening, speaking, reading, writing and grammar at COLT, when the assessment studies herein were conducted, followed a process-based approach which required process-based assessment to match what the students' practice. Henceforth, the studies grouped in Cluster 1 and parts of Cluster 2 show criteria for constructing high-quality language skill tests that measure both the process and the product of each skill. These studies outline the essential characteristics of effective assessment, such as comprehensive coverage of instructional content, integration of process- and product subskills, alignment with course objectives, and the use of tasks that require comprehension and production, analysis, application, synthesis, inference, and higher-order thinking. The remaining clusters, particularly the experimental studies, the multi-skill assessment study, and the studies on diagnostic morphology, spelling, and technology-enhanced assessment, apply these criteria in practice. The listening, pronunciation, speaking, reading, writing, vocabulary, grammar, morphology, and spelling tests used in these studies were designed in accordance with the principles established in the earlier clusters. In many cases, the final course exam served as the post-test, and these exams themselves embodied the same standards of comprehensiveness, cognitive demand, and process-product integration. This alignment demonstrates that the assessment procedures across the corpus are grounded in a unified theoretical and pedagogical framework, ensuring consistency in how language skills are conceptualized, measured, and interpreted throughout the author's research program.

4. Discussion

Findings of this SR reveal a coherent and sustained research trajectory in the assessment of language skills across a corpus of 40 studies spanning nearly two decades. Collectively, the studies demonstrate how listening, pronunciation, speaking, reading, writing, vocabulary, grammar, morphology, spelling, and research skills have been defined, operationalized, and assessed within varied instructional contexts. A central theme across the corpus is the author's consistent emphasis on aligning assessment practices with instructional goals, assigned textbook instructional approach, learner needs, and the cognitive and linguistic processes underlying each skill. Studies in Cluster 1 and parts of Cluster 2 articulate the foundational criteria for constructing high-quality language skill tests—criteria that emphasize comprehensive content coverage, integration of process- and product-based skills, alignment with course objectives, and the use of tasks that require inference, synthesis, and higher-order thinking. These principles are reflected in the empirical, diagnostic, and experimental studies across Clusters 3–7, where the listening, speaking, reading, writing, vocabulary, grammar, morphology, and spelling tests used in classroom-based and technology-enhanced interventions were designed in accordance with these criteria.

Across the seven thematic clusters, the studies illustrate both convergence and diversity in assessment approaches. While all studies share a principled orientation toward sound assessment design, they differ in methodological focus—ranging from diagnostic analyses of learner errors, to experimental evaluations of technology-enhanced instruction, to the development of rubrics, tools, and task-based procedures. This diversity enriches the overall understanding of language skill assessment and underscores the need for flexible, context-sensitive frameworks that accommodate differences in learner proficiency, instructional settings, task types, and pedagogical goals. The corpus also demonstrates how assessment practices evolve in response to technological advancements, shifts in instructional modalities, and changing learner needs. By integrating conceptual principles with practical assessment models, detailed rubrics, and empirically validated testing procedures, the author's work contributes to bridging the gap between theory and practice and offers adaptable frameworks for assessing a wide range of language skills in EFL contexts.

4.1 *A Process–Product Perspective in Language Skill Assessment*

A useful interpretive note for understanding the findings of this SR is the distinction between process and product in language-skill development. In this framework, any observable linguistic performance, whether in listening, speaking, reading, writing, vocabulary, grammar, morphology, spelling, and research, represents a final product that is preceded by a sequence of cognitive, linguistic, and strategic processes. These processes may include decoding, morphological analysis, synthesizing information, monitoring comprehension, planning, predicting, and revising. The reviewed studies collectively suggest that effective assessment must take into consideration both dimensions: the visible end product and the underlying processes that generate it. This theoretical perspective provides a foundation for interpreting the assessment practices and findings synthesized in this SR.

4.2 *Meta Conclusion*

The meta-level synthesis of the 40 studies reveals a coherent and sustained research program that advances a comprehensive understanding of language-skill assessment across multiple skill areas. Collectively, the studies demonstrate that effective assessment in EFL contexts requires more than the construction of test items; it demands a principled, skill-specific, and pedagogically grounded approach that aligns assessment practices with instructional goals, learner needs, and the linguistic and cognitive demands of each skill. Across listening, speaking, reading, writing, vocabulary, grammar, morphology, spelling and research-related skills, the author consistently emphasizes the importance of transparent assessment criteria, detailed process and product skill decomposition, and comprehensive content coverage.

A central conclusion is that assessment functions most effectively when it is embedded within instruction and used as a tool for diagnosis, feedback, and skill development rather than as a purely pass/fail result. The studies repeatedly show that well-designed assessments, supported by tables of specifications, clear instructions, appropriate difficulty levels, and valid scoring procedures, enhance learners' performance and provide instructors with actionable insights. The corpus also highlights the value of technology-enhanced tools, such as digital rubrics and mobile apps, in promoting consistency, transparency, and learner engagement.

Taken together, the studies form a unified framework that positions language-skill assessment as a dynamic, context-sensitive, and pedagogically integrated process. This framework underscores the need for assessments that are reliable, valid, discriminative, and aligned with real instructional goals and practices. The meta-conclusion affirms that the author's research program contributes a robust, practice-oriented model for understanding and improving language-skill assessment in EFL settings.

4.3 *Meta Interpretation*

A meta-interpretive reading of the 40-study corpus reveals an underlying conceptual stance that positions language-skill assessment as an integrated, developmental, and context-responsive process rather than a static measurement activity. Across the author's studies, assessment is consistently interpreted as a pedagogical tool that shapes learning, diagnoses weaknesses, and guides instructional decision-making. This is evident in the repeated emphasis on detailed process and product skill decomposition, explicit assessment criteria, and the alignment of test content with instructional objectives and content covered in the textbooks and in the classroom. The corpus collectively suggests that language skills, whether listening, speaking, reading, writing, vocabulary, grammar, morphology, spelling or research skills, are multidimensional constructs that require equally multidimensional assessment procedures.

At a deeper level, the studies interpret assessment as a form of a structured mechanism that supports learners' progression from controlled tasks to more complex, authentic performance. The author's consistent use of tables of specifications, analytic rubrics, and task-based evaluation models reflects an interpretive belief that transparency and structure empower learners and reduce ambiguity in performance expectations. The integration of technology, digital rubrics, mobile apps, online platforms, MP3-based speaking tests, further reinforces an interpretive stance that assessment should evolve alongside pedagogical and technological shifts, ensuring fairness, consistency, and accessibility.

The corpus also conveys an implicit teacher's role in assessment quality which is part of instructional quality. Studies in Cluster 4 show that instructor expertise, pedagogical choices, and feedback practices significantly shape students' performance, suggesting that assessment cannot be meaningfully separated from the broader instructional environment. This interpretation positions assessment as a relational practice, influenced by teacher knowledge, learner characteristics, course design, and contextual constraints.

Taken together, the meta-interpretation reveals that the author's research program advances a holistic view of language-skill assessment: one that is diagnostic rather than punitive, developmental rather than static, and deeply embedded in the realities of EFL teaching and learning. This interpretive stance underscores the need for assessment systems that are principled, transparent, skill-specific, and adaptable to diverse educational contexts.

4.4 *Cross Cutting Insights*

A cross-cluster analysis of the 40 studies reveals several insights that cut across skill types, assessment tools, and instructional contexts. First, the corpus consistently underscores the centrality of alignment between instructional objectives, skill-specific demands, test content, and scoring procedures. Whether assessing listening, speaking, reading, writing, vocabulary, grammar, morphology, spelling or research skills, the studies emphasize that assessment must directly reflect what learners are taught and what they are expected to perform in authentic academic tasks.

Second, the studies collectively highlight the importance of skill decomposition into process and product subskills as a foundation for valid assessment. Across clusters, the author breaks down complex language skills into fine-grained subskills—word-level, sentence-level, phonological, morphological, syntactic, semantic, discourse-level, and task-based components. This decomposition

not only guides test construction but also supports diagnostic feedback, enabling instructors to identify specific areas of weakness and design targeted remediation.

Third, the corpus demonstrates a persistent commitment to transparency and structure in assessment. Tables of specifications, comprehensiveness, analytic rubrics, explicit instructions, and clear scoring criteria appear repeatedly across the studies. These tools serve not only to enhance reliability and fairness but also to make assessment processes visible and understandable to learners, thereby reducing ambiguity and supporting learner autonomy.

Fourth, the studies reveal a strong orientation toward technology-mediated assessment. Digital rubrics, mobile flashcard apps, online platforms, and MP3-based speaking tests are used to enhance efficiency, consistency, and accessibility. Technology is not treated as an add-on but as an integral component of modern assessment practice.

Finally, the corpus highlights the inseparability of assessment and pedagogy. Findings from Cluster 4 show that instructor expertise, pedagogical choices, and feedback practices significantly influence learners' performance. This suggests that assessment quality is deeply embedded in the broader instructional ecosystem and cannot be isolated from teaching methods, course design, or contextual factors. Together, these cross-cutting insights reveal a unified vision of language-skill assessment as principled, transparent, diagnostic, and pedagogically integrated, matching learner needs, instructional realities, and technological advancements.

4.5 *Implications*

The findings of this SR carry several important implications for language-skill assessment in education contexts. First, the strong emphasis on skill decomposition into process and product subskills across the reviewed studies indicates that instructors should adopt structured, principled approaches to evaluating learners' performance. Breaking complex skills, such as speaking, reading, vocabulary, or writing, into measurable subskills enables more diagnostic assessment and allows teachers to identify specific areas of weakness.

Second, the studies highlight the pedagogical value of integrating assessment into the learning process rather than treating it as a purely summative event. Assessments supported by clear instructions, analytic rubrics, and tables of specifications promote transparency and help learners understand expectations, leading to more intentional and strategic learning. This suggests that institutions should encourage formative assessment practices and embed feedback within regular instruction.

Third, the widespread use of technology-enhanced assessment tools across the corpus demonstrates that digital assessment is now essential. Online rubrics, mobile applications, and automated scoring systems improve objectivity, fairness, consistency, efficiency, and accessibility. Teacher-training programs should therefore prioritize digital assessment literacy to ensure ethical and effective implementation.

Fourth, the review underscores the central role of teacher expertise in shaping assessment quality. Variations in pedagogical choices, feedback practices, and familiarity with assessment principles directly influence student outcomes. This highlights the need for sustained professional development in test design, rubric construction, and data-driven decision-making.

Finally, the findings must be interpreted within the broader instructional context of the program at COLT. The Interactions I & II and Mosaic I & II listening, speaking, reading, writing and grammar textbooks assigned by COLT, are based on a process approach to skill development. Within this approach, every skill has both a process dimension (planning, predicting, monitoring, decoding, synthesizing) and a product dimension (the final performance). The studies in this assessment corpus consistently reflect this duality, emphasizing that students must be explicitly taught the process skills in order to produce strong linguistic products. This has direct implications for tests and rubrics, which should capture not only the final output but also the cognitive and linguistic operations that lead to it.

Collectively, these implications point toward a more integrated, transparent, and pedagogically aligned approach to language-skill assessment—one that supports both teaching and learning in meaningful and sustainable ways.

4.6 *Positioning This SR Within the Global Language Skill Assessment SR/MA Research*

Within the broader scope of global SRs and MAs on language-skill assessment, this SR occupies a distinctive position by synthesizing a unified, author-bounded corpus that spans 20 years of continuous research. While international SR/MA studies typically aggregate findings from multiple researchers, diverse contexts, and heterogeneous methodologies, the present SR offers a rare opportunity for tracing the evolution of a single, coherent research program focused on the principled assessment of language skills in EFL settings. This unique design allows for a level of conceptual, methodological, and pedagogical continuity that is often absent in large-scale global syntheses.

Globally, SR/MA research on language assessment tended to emphasize broad themes such as test validity, reliability, washback effects, technology-enhanced assessment, and skill-specific measurement practices. The current SR aligns with these international trends but contributes a more granular, practice-oriented perspective by documenting how assessment principles are operationalized across reading, vocabulary, speaking, writing, grammar, listening, and research-related skills. Unlike many global reviews that focus on standardized testing or large-scale assessment systems, this SR foregrounds classroom-embedded assessment practices, diagnostic uses of testing, and the integration of digital tools such as rubrics and mobile applications.

Furthermore, this SR expands the global conversation by highlighting assessment challenges and innovations within EFL contexts that are underrepresented in international SR/MA literature, particularly the assessment of ESP reading, research skills, and skill-specific rubrics in higher education. By synthesizing a corpus that systematically links assessment design, instructional alignment, and learner needs, the review offers a model that complements and enriches global research traditions. It demonstrates how sustained, context-sensitive inquiry can generate insights that are both locally grounded and globally relevant.

In this sense, the present SR not only situates itself within the international body of language-assessment research but also contributes a distinctive, practice-driven framework that can inform future SR/MA studies seeking to integrate pedagogical, technological, and contextual dimensions of skill assessment.

4.7 *Positioning This Language Skill Assessment SR Within Process–Product Theories*

This SR can be positioned within the broader distinction between process and product theories in language-skill development which argue that observable linguistic performance (the product) is the culmination of a series of cognitive, linguistic, and strategic operations (the process). Each language skill, listening, speaking, reading, writing, grammar, and vocabulary, has its own set of process subskills that ultimately shape the quality of the final product. The studies included in this SR consistently highlight the importance of skill decomposition, strategy instruction, and diagnostic assessment, all of which align closely with process-oriented models of learning.

The author's long-term teaching experience further reinforces this theoretical positioning. Instruction and assessment practices that emphasize morphological analysis, rapid lexical processing, cross-course skill transfer, and higher-order thinking reflect a process-based pedagogy that prepares learners for real-world academic and professional demands. Students' long-term retention of vocabulary, their ability to apply analytical strategies across courses, and their improved performance in advanced reading and translation courses provide empirical support for the value of process-oriented instruction and assessment. These outcomes demonstrate that when learners are taught the underlying processes—not just the final products—they develop deeper, more durable, and more transferable linguistic competence.

Thus, this SR not only synthesizes existing research but also situates language-skill assessment within a theoretical framework that recognizes the centrality of cognitive processes in shaping linguistic performance. This positioning strengthens the contribution of the review by linking empirical findings to a well-established body of theory in applied linguistics and educational assessment.

4.8 *Integrating Practice-Based Test Construction Within Process–Product Theories*

A key feature of this research program, and one that aligns directly with process–product theories, is the reciprocal relationship between instructional practice and test construction. The assessment criteria used across the corpus were not developed in abstraction; rather, they emerged from long-term, practice-based test development conducted in workshops, teacher-training sessions, and classroom settings. The tests used in the experimental studies and those in Cluster 2 were built on the same principles demonstrated in these workshops, and the workshops themselves were informed by the performance patterns observed in actual test administrations (See Cluster 1 studies). This bidirectional flow between practice and theory exemplifies the process–product model: instructional processes shaped the design of assessment products, and the performance on these products, in turn, refined the instructional processes.

Moreover, the posttests used in the experimental studies were repeatedly administered to different cohorts over multiple semesters, consistently demonstrating strong discriminating power, comprehensive coverage, appropriate difficulty level, reliability, and validity. Institutional policies, such as not returning final exam answer sheets to students, maintaining a minimum four-month interval between administrations, and using parallel versions of final exams, ensured that test content remained secure and that repeated use did not compromise fairness or validity. This stability allowed the tests to function as reliable indicators of the cognitive and linguistic processes emphasized in instruction, reinforcing the process–product alignment that underpins the entire research program.

4.9 *Limitations of This SR*

Although the studies synthesized in this SR are methodologically rigorous and grounded in a long-term, practice-based assessment tradition, three limitations should be acknowledged. First, all studies were conducted within a single institutional context. This institutional coherence provides a rare advantage, ensuring stable curricula, consistent assessment practices, and longitudinal comparability, but it also means that the findings reflect the pedagogical and assessment culture of that specific environment. As such, the applicability of the results to institutions with different instructional models or learner populations, even within Saudi Arabia and even by different instructors, may be naturally limited.

Second, 31 of the 40 studies were conducted with Level 1 students enrolled in Listening I, Speaking I, Reading I, Writing I, Vocabulary I, and Grammar I. These courses provide rich opportunities for examining foundational language development, but they do not represent the full range of advanced skills taught in Levels 2, 3, and 4. The college offers four levels of Listening, Speaking, Reading, and Writing, two levels of Vocabulary, and three levels of Grammar, yet the published studies focus primarily on Level 1 skill acquisition. As a result, the research program offers strong empirical coverage of beginning-level learning but does not systematically examine how assessment principles operate with higher-level learners who engage with more complex grammatical structures, advanced reading and writing tasks, or higher-order listening and speaking skills. Furthermore, the studies were conducted with the author's own students, whom she taught and tested. She could not test students in other levels or sections taught by other colleagues due to validity and reliability issues, intervening variables and administrative restrictions.

Third, the specific subskills examined in the published studies do not represent the full range of process skills taught in the courses. For example, grammar and vocabulary instruction typically covers complex sentences, passive voice, reported speech, collocations, tenses, question formation, lexical meaning, synonyms, acronyms, and others. However, these areas were not the focus of standalone research studies. Instead, the author often selected a single subskill—such as plural formation or adjective-forming suffixes—from students' responses on a final exam or weekly quiz and analyzed it in depth. It is noteworthy to say that all subskills within each course were consistently assessed through weekly quizzes and final examinations, even if they did not appear as individual research topics in the published corpus. As a result, the studies highlight selected skill components rather than providing an extensive coverage of every subskill taught across the broader curriculum.

4.10 *Future Research Directions*

The synthesis of the 40 studies highlights several promising directions for future research in language-skill assessment. First, future studies could explore how the assessment principles established at Level 1 operate with learners in higher-level courses (Levels 2, 3, and 4), where students engage with more advanced reading, writing, listening, speaking, grammar, and vocabulary skills. Investigating these upper-level contexts would provide a broader picture of skill development across the full curriculum and allow for comparisons between foundational and advanced proficiency stages. Second, future research may focus on specific grammatical and lexical subskills that were regularly assessed in instructional practice—such as complex sentence structures, passive voice, reported speech, collocations, and tense systems—but were not the primary focus of published studies. Dedicated empirical investigations of these subskills could offer finer-grained insights into learners' developmental trajectories and the effectiveness of targeted assessment tools. Third, expanding the research program to include additional instructional contexts—such as different institutions, parallel programs, or varied learner populations—would allow for examining the transferability of the assessment principles developed within the original institutional setting. Such comparative work could illuminate how contextual factors shape assessment outcomes and whether the process-based framework retains its effectiveness across diverse environments. Finally, future studies may incorporate further technological tools or digital assessment platforms to explore how emerging technologies can support or enhance process-oriented assessment practices. This line of inquiry would align the research program with current developments in language testing and provide opportunities for innovation in test design, scoring, and feedback.

5. Recommendations

Based on the synthesis of the 21 studies, several practical recommendations can be offered to instructors, curriculum designers, and assessment specialists. First, instructors are encouraged to adopt skill-specific assessment frameworks that clearly articulate the subskills to be measured and align test content with instructional objectives. Detailed tables of specifications, explicit instructions, and transparent scoring criteria should be standard components of all assessment practices. Skill-specific assessment should focus on both the listening, speaking, reading, writing, vocabulary, grammar skills process and product especially when the target students are trained to be translators and interpreters.

Second, teachers should integrate diagnostic and formative assessment into their regular instructional routines. Tests should not be limited to summative evaluation but should be used to identify learners' strengths and weaknesses, guide remediation, and

support ongoing skill development. Providing students with constructive feedback, supported by rubrics and sample responses—can significantly enhance learning outcomes.

Third, institutions should invest in assessment literacy training for instructors. The findings show that teacher expertise, pedagogical choices, and familiarity with assessment principles directly influence student performance. Professional development programs focusing on test design, rubric construction, and technology-enhanced assessment tools can improve the quality and fairness of classroom evaluation.

Fourth, instructors are encouraged to make greater use of digital assessment tools, such as iRubrics, mobile learning applications, and online platforms. These tools enhance transparency, consistency, and efficiency, and they support student engagement and self-assessment. Finally, curriculum designers should ensure that assessment practices reflect the cognitive and linguistic complexity of academic tasks, particularly in ESP and higher-education contexts, where learners must navigate specialized vocabulary, disciplinary texts, and research-related skills.

6. Conclusion

This SR synthesizes a sixteen-year research program that offers a comprehensive, practice-oriented framework for assessing language skills. Across 21 studies, the author demonstrates a consistent commitment to principled assessment design, skill-specific measurement focusing on the process and product aspects of each skill, and pedagogically grounded evaluation practices. The corpus highlights the importance of aligning assessment with instruction, decomposing complex skills into measurable components, and using assessment as a tool for diagnosis, feedback, and learning—not merely as a mechanism for grading.

The SR also underscores the growing role of technology in enhancing transparency, reliability, and learner engagement. Digital rubrics, mobile applications, and online assessment platforms such as Blackboard emerge as powerful tools that support both instructors and students. At the same time, the findings reveal that assessment quality is deeply intertwined with teacher expertise, pedagogical choices, and contextual factors, emphasizing the need for ongoing professional development and institutional support.

Overall, this SR contributes a coherent and context-sensitive model of language-skill assessment that focuses on both the skill process and product and that is both locally grounded and globally relevant. It provides a foundation for future research, encourages more rigorous and diversified assessment practices, and offers practical guidance for improving the evaluation of language skills in EFL settings. Through its integrated perspective, the review affirms that effective assessment is not an isolated act but a dynamic, evolving process that shapes, and is shaped by, the broader ecosystem of teaching and learning.

Conflicts of Interest: The author declares no conflict of interest.

ORCID ID: https://orcid.org/0000-0002-6255-1305

Publisher's Note: All claims expressed in this article are solely those of the author and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Al-Jarf, R. (2026a). An integrative review of studies on teaching English for art education purposes to Ph.D. students. *International Journal of Arts and Humanities Studies*, *6*(2), 01-15. DOI: 10.32996/ijahs.2026.6.2.1. Google Scholar

[2] Al-Jarf, R. (2026b). An interpretive systematic review of a researcher's contributions to EFL reading instruction: Themes, methods, and pedagogical insights. *Journal of English Language Teaching and Applied Linguistics*, *8*(4), 01-22. https://doi.org/10.32996/jeltal.2026.8.4.1. Google Scholar

[3] Al-Jarf, R. (2026c). An Integrative Systematic review of Studies Across Diverse Educational Evaluation Domains. *British Journal of Teacher Education and Pedagogy*, *5*(3), 40-60. https://doi.org/10.32996/bjtep.2026.5.3.5. Google Scholar

[4] Al-Jarf, R. (2026d). A self systematic review of translation error studies (2000–2025): The Case of students' errors in English–Arabic and Arabic–English translation. *International Journal of Translation and Interpretation Studies*, *6*(1), 16-32. DOI: 10.32996/ijtis.2026.6.1.2. Google Scholar

[5] Al-Jarf, R. (2026e). A self-systematic review of mobile apps for developing multiple language skills in EFL. *Journal of Computer Science and Technology Studies*, *8*(3), 14-29. DOI: 10.32996/jcsts.2026.8.3.2. Google Scholar

[6] Al-Jarf, R. (2026f). A systematic review of Studies on Adult Reading Practices, Interests, Habits and Challenges. *Journal of Humanities and Social Sciences Studies*, *8*(3), 114-129. https://doi.org/10.32996/jhsss.2026.8.3.9. Google Scholar

[7] Al-Jarf, R. (2026g). A systematic review of studies on pronunciation instruction and practice in L2 (2005-2025). *Journal of English Language Teaching and Applied Linguistics*, *8*(1), 10-26. DOI: 10.32996/jeltal.2026.8.1.2. Google Scholar

[8] Al-Jarf, R. (2026h). A systematic review of studies on teaching reading in Arabic to grades 1–12: Textbooks, skills, and learning outcomes. *Journal of Learning and Development Studies*, *6*(5), 01-19. https://doi.org/10.32996/jlds.2026.6.5.1. Google Scholar

[9] Al-Jarf, R. (2026i). Arabic–English transliteration of personal names and public signages: A Systematic review and Meta-analysis. *British Journal of Applied Linguistics*, *6*(1), 01-14. DOI: 10.32996/bjal.2025.6.1.1. Google Scholar

[10] Al-Jarf, R. (2026j). Children's language acquisition and development in Saudi Arabia: A Systematic review and meta analysis. *Journal of Learning and Development Studies*, *6*(1), 18-37. DOI: 10.32996/jlds.2026.6.1.3. Google Scholar

[11] Al-Jarf, R. (2026k). Classroom practices, writing enhancement and creativity among EFL struggling students: A systematic review. *Journal of World Englishes and Educational Practices*, *8*(1), 20-38. DOI: 10.32996/jweep.2026.8.1.3. Google Scholar

[12] Al-Jarf, R. (2026l). Collaborative learning and teaching in digital environments: A systematic review of two decades of research. *Journal of Computer Science and Technology Studies*, *8*(4), 25-40. https://doi.org/10.32996/jcsts.2026.8.4.2 Google Scholar

[13] Al-Jarf, R. (2026m). Effectiveness of mind-mapping on multiple English language skills in the Saudi Context: A systematic review. *Frontiers in English Language and Linguistics*, *3*(1), 01-10. DOI: 10.32996/fell.2026.3.1.1. Google Scholar

[14] Al-Jarf, R. (2026n). Inadequate staffing and large class sizes in Saudi EFL and translation programs: An integrative analysis of empirical studies. *British Journal of Teacher Education and Pedagogy*, *5*(1), 19-27. DOI: 10.32996/bjtep.2026.5.1.3. Google Scholar

[15] Al-Jarf, R. (2026o). Innovative word formation and pluralization processes in Arabic: A systematic review. *Journal of Humanities and Social Sciences Studies*, *8*(1), 44-60. DOI: 10.32996/jhsss.2026.8.1.6. Google Scholar

[16] Al-Jarf, R. (2026p). Systematic review and meta-analysis of 2024–2025 studies on AI Arabic translation, linguistics and pedagogy. *Frontiers in Computer Science and Artificial Intelligence*, *5*(1), 07-27. DOI: 10.32996/jcsts.2026.5.1.2. Google Scholar

[17] Al-Jarf, R. (2026q). Three decades of ESP Innovation: A review of research across specialized and underexplored domains. *British Journal of Teacher Education and Pedagogy*, *5*(2), 19-31. DOI: 10.32996/bjtep.2026.5.2.3. Google Scholar

[18] Al-Jarf, R. (2025). Can ESL students identify emphatic features of advertisements? *Journal of World Englishes and Educational Practices* , *7*(2), 01-11. https://doi.org/10.32996/jweep.2025.7.2.1. Google Scholar

[19] Al-Jarf, R. (2024). Expressions of impossibility in Arabic and English: unveiling students' translation difficulties. *International Journal of Linguistics, Literature and Translation*, *7*(5), 68-76. DOI: 10.32996/ijllt.2024.7.5.9. ERIC ED651472. Google Scholar

[20] Al-Jarf, R. (2023a). Challenges faced by Arab peer-reviewers. *International Journal of Arts and Humanities Studies*, *3*(4), 31-41. Google Scholar

[21] Al-Jarf, R. (2023b). Equivalence problems in translating ibn (son) and bint (daughter) fixed expressions to Arabic and English. *International Journal of Translation and Interpretation Studies, 3, 2, 1-15.* DOI: 10.32996/ijtis.2023.3.2.1. ERIC ED628181 Google Scholar

[22] Al-Jarf, R. (2023c). Problems of identifying lexical and syntactic features of legal documents by undergraduate EFL students. *Journal of Pragmatics and Discourse Analysis, 1*(1), 31-39. DOI: 10.32996/jpda.2023.2.1.3. ERIC ED627248. Goolge Scholar

[23] Al-Jarf, R. (2023d). Numeral-based English and Arabic Formulaic Expressions: Cultural, Linguistic and Translation Issues. *British Journal of Applied Linguistics, 3,1,* 25-34. https://doi.org/10.32996/bjal.2023.3.1.2. ERIC ED628151. Google Scholar

[24] Al-Jarf, R. (2023e). Testing multiple vocabulary associations for effective long term learning. British Journal of Teacher Education and Pedagogy, 2(3), 57-71. DOI: 10.32996/bjtep.2023.2.3.6. ERIC ED634388. Google Scholar

[25] Al-Jarf, R. (2023f). Time metaphors in English and Arabic: translation challenges. International Journal of Translation and Interpretation Studies (IJTIS), 3, 4, 68-81 https://doi.org/10.32996/ijtis.2023.3.4.8. Google Scholar

[26] Al-Jarf, R. (2022a). Arabic and English dar (house) and bayt (home) expressions: Linguistic, translation and cultural issues. Journal of Pragmatics and Discourse Analysis (JPDA), 1(1), 1-13. ERIC ED624367 Google Scholar

[27] Al-Jarf, R. (2022b). Deviant Arabic transliterations of foreign shop names in Saudi Arabia and decoding problems among shoppers. International Journal of Asian and African Studies (IJAAS), 1(1), 17-30. DOI: 10.32996/ijaas.2022.1.1.3. Google Scholar

[28] Al-Jarf, R. (2022c). Difficulties in learning English plural formation by EFL college students. International Journal of Linguistics, Literature and Translation (IJLLT), 5(6), 111-121. Doi:10.32996/ijllt.2022.5.6.13. ERIC ED620200. Google Scholar.

[29] Al-Jarf, R. (2022d). Grade inflation in language and translation courses at Saudi schools and universities. *British Journal of Teacher Education and Pedagogy*, *1*(2), 08-25. Google Scholar

[30] Al-Jarf, R. (2022e). Grade inflation at Saudi universities before, during and after the pandemic: A comparative study. *Journal of Humanities and Social Sciences Studies (JHSSS), 4*(4). 111-125. DOI: 10.32996/jhsss.2022.4.4.15. ERIC ED623003. Google Scholar

[31] Al-Jarf, R. (2022f). Issues in translating English and Arabic Common names of chemical compounds by student-translators in Saudi Arabia. In Kate Isaeva (Ed.). Special Knowledge Mediation: Ontological & Metaphorical Modelling. Springer. DOI: 10.1007/978-3-030-95104-7. Google Scholar

[32] Al-Jarf, R. (2022g). MA and Ph.D. thesis evaluation at Saudi universities: Problems and solutions. *Eurasian Arabic Studies, 5*(2), 88–106. DOI: 10.26907/2619-1261.2022.5.2.88-106. Google Scholar

[33] Al-Jarf, R. (2022h). Online exams in language, linguistics and translation courses during the pandemic in Saudi *Arabia. Journal of World Englishes and Educational Practices (JWEEP), 4*(3), 14-25. DOI: 10.32996/jweep.2022.4.3.2. ERIC ED622401. Google Scholar

[34] Al-Jarf, R. (2022i). Role of instructor qualifications, assessment and pedagogical practices in EFL students' grammar and writing proficiency. *Journal of World Englishes and Educational Practices (JWEEP), 4*(1), 18-33. DOI: 10.32996/jweep.2022.4.2.2. ERIC ED618315. Google Scholar

[35] Al-Jarf, R. (2021). Collaborative mobile ebook reading for struggling EFL college readers. IOSR Journal of Research and Methods in Education. 11, 6, 32-42. DOI: 10.9790/7388-1106023242. ERIC ED618023. Google Scholar

[36] Al-Jarf, R. (2021a). Critical analysis of translation tests in 18 specialized translation courses: Shortcomings and recommendations. *EJ-EDU-European Journal of Education and Pedagogy (ej-edu.org), 3*(5). Google Scholar

[37] Al-Jarf, R. (2021b). EFL female college students and instructors' preferred method of speaking assessment: A perspective from Saudi Arabia. *Asian Journal of Education and Social Studies (AJESS), 16*(3), 38-50. doi: 10.9734/ajess/2021/v16i330403. Google Scholar

[38] Al-Jarf, R. (2021c). EFL Students' difficulties with lexical and syntactic features of news headlines and news stories. Technium Social Sciences Journal, 17(1), 524–537. **ERIC** ED618106. Google Scholar

[39] Al-Jarf, R. (2021d). How much material do EFL college instructors cover in reading courses? *Journal of Applied Linguistics and Language Research (JALLR), 8*(1), 65-79. ERIC ED620414. Google Scholar

[40] Al-Jarf, R. (2021e). Standardized test preparation with mobile flashcard apps. *United International Journal for Research & Technology (UIJRT)*, 3(2), 33-40. ERIC ED616917. Google Scholar

[41] Al-Jarf, R. (2021f). Testing reading for specific purposes in an art education course for graduate students in Saudi Arabia. *International Journal of Advance and Innovative Research,* 8 (1), 32-42. Google Scholar

[42] Al-Jarf, R. (2021g). *Testing reading for specific purposes in an art education course for graduate students in Saudi Arabia*. International Conference on Research and Development in Science, Technology and Management in the Current Era. Indian Academicians and Researchers Association (IARA), India. February 21. Google Scholar

[43] Al-Jarf, R. & Mingazova, N. (2020a). *Evaluation of Russian Arabic language teaching textbooks in the light of CEFR criteria*. ARPHA Proceedings #3. Pp. 101-129. VI International Forum on Teacher Education, Kazan Federal University, Russia. DOI: 10.3897/ap.2.e0101. ERIC ED613172. https://ap.pensoft.net/article/22255. Google Scholar

[44] Al-Jarf, R. (2020b). How EFL college instructors can create and use grammar iRubrics. *Journal of Global Research in Education and Social Science (JOGRESS), 14*(3): 22-38. Google Scholar

[45] Al-Jarf, R. (2019a). EFL freshman students' difficulties with phoneme-grapheme relationships. 5th VietTESOL International Convention. Hue University of Foreign Languages, Hue, Vietnam. October 11-12. Google Scholar

[46] Al-Jarf, R. (2019b). Freshman students' difficulties with English adjective-forming suffixes. *Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL), 6*(1), 169-180. Google Scholar

[47] Al-Jarf, R. (2019c). Teaching reading to EFL Arabic students online. Eurasian Arabic Studies, 8, 57-75. ERIC ED613084. Google Scholar

[48] Al-Jarf, R. (2019d). Translation students' difficulties with English and Arabic color-based metaphorical expressions. Fachsprache, 41 (Sp. Issue), 101-118. Doi: 10.24989/fs.v41iS1.1774. ERIC ED622935. Google Scholar

[49] Al-Jarf, R. (2018a). Effect of background knowledge on auditory comprehension in interpreting courses. In Renata Jancarikova (Ed.) Interpretation of Meaning across Discourse, pp. 97-108. Muni Press, Brno, Czech Republic. https://doi.org/10.5817/CZ.MUNI.M210-8947-2018. Google Scholar

[50] Al-Jarf, R. (2018b). Effect of background knowledge on auditory comprehension in interpreting courses. 4th Brno Conference on Linguistics Studies in English titled "Interpretation of Meaning in Spoken and Written Discourse". Masaryk University, Brno, Czech Republic. Sept 16-17, 2010. ERIC ED665097. Google Scholar

[51] Al-Jarf, R. (2018c). First, second and third grade students' word identification difficulties. *Eurasian Arabic Studies*, 8, 22-93. Google Scholar

[52] Al-Jarf, R. (2017a). Issues in translating Arabic om- and abu-expressions. *Alatoo Academic Studies, 3,* 278-282. ERIC ED613247. Google Scholar

[53] Al-Jarf, R. (2017b). *What teachers should know about reading tests*. 3rd ELT Conference. Ibri College of Technology, Oman. April 5. Google Scholar

[54] Al-Jarf (2016). Translation of English and Arabic binomials by advanced and novice student translators. In Larisa Ilynska and Marina Platonova (Eds) Meaning in Translation: Illusion of Precision (Pp. 281-298). Cambridge Scholars Publishing. ERIC ED639264. Google Scholar

[55] Al-Jarf, R. (2015a). *Assessing EFL college instructors' performance with digital rubrics*. In *Teaching and learning in Saudi Arabia: Perspectives from higher education* (pp. 1-30). Rotterdam: Sense Publishers. Google Scholar

[56] Al-Jarf, R. (2015b). *Issues in assessing the speaking skill in EFL*. international conference on language testing and assessment. Guangzhou, China. November 27-30. Google Scholar

[57] Al-Jarf, R. (2015c). *Textbook evaluation checklist*. https://www.researchgate.net/profile/Reima-Al-Jarf/publication/280943540. Google Scholar

[58] Al-Jarf, R. (2015d). *What teachers should know about vocabulary tests for EFL freshman students*. International Conference on Language Testing and Assessment. Guangzhou, China. November 27-30. Google Scholar www.researchgate.net/publication/352351036. Google Scholar

[59] Al-Jarf, R. (2014). Collaborative mobile ebook reading by translation students. September 18-19. 7th International Conference on the Importance of Learning Foreign Languages. University of Maribor. September 11-12. Google Scholar

[60] Al-Jarf, R. (2013a). *Assessing graduate students' research skills in EFL*. ESP Conference 2013 entitled "Assessing Graduate Students' Research Skills in English". University of Niš, Serbia. May 17-19. Google Scholar

[61] Al-Jarf, R. (2013b). *Enhancing freshman students' performance with online reading and writing activities*. 9th eLearning and Software for Education Conference (eLSE). Bucharest, Romania. 2, 524-530. DOI: 10.12753/2066-026X-13-193. ERIC ED623858. Google Scholar

[62] Al-Jarf, R. (2012a). *Creating and sharing vocabulary iRubrics*. 11th Asia CALL Conference. Ho Chi Minh City Open University, Vietnam, Nov. 16 – 18. Google Scholar

[63] Al-Jarf, R. (2012b). *Mobile technology and student autonomy in oral skill acquisition*. In Javier E. Díaz Vera's Left to My Own Devices: Learner Autonomy and Mobile-Assisted Language Learning, Innovation and Leadership in English Language Teaching, 6, 105-129. Brill. Doi: 10.1163/9781780526478_007. Google Scholar

[64] Al-Jarf, R. (2011a). Creating and sharing writing iRubrics. In Paul Robertson and Roger Nunn (Eds.), *the Asian EFL Journal Professional Teaching Articles* – CEBU Issue 51, April, 41-62. ERIC ED638501. Google Scholar

[65] Al-Jarf, R. (2011b). *Developing and testing reading skills through art texts*. In S.V. Lobanov, S. Bulaeva, S. Somova, N.P. Chepel (Eds), Language and Communication through Culture. 168-176. Ryazan State University. Russia. Google Scholar

[66] Al-Jarf, R. (2011c). *Empowering EFL teachers and students with grammar iRubrics*. Proceedings of the Eleventh Annual ELT Conference entitled: "Empowering Teachers and Learners". Sultan Qaboos University, Oman. Pp. 50-66. ERIC ED638284. https://doi.org/10.2139/ssrn.3851495 . Google Scholar

[67] Al-Jarf, R. (2011). *Mobile technology and student autonomy in oral skill acquisition*. International Conference on Mobile Learning and Autonomy in Second Language Acquisition. Toledo, Spain, Sep 17-19. ERIC ED638511.

[68] Al-Jarf, R. (2010a). *Creating and Sharing iRubrics Using RCampus*. 15th TCC Online Conference "Yesterday, Today & Tomorrow: Communication, Collaboration, Communities, Mobility and Best Choices". April 20-22. Google Scholar

[69] Al-Jarf, R. (2010b). Integrating RCampus in College Reading and Writing for Translation students. Touchpoint 2010 International Conference on Technology in Education. Manila, Philippines, March 5-6. ERIC ED609048. Google Scholar

[70] Al-Jarf, R. (2010c). Translation students' difficulties with English neologisms. Analele Universităţii "Dunărea De Jos" Din Galaţi Fascicula XXIV ANUL III (2). 431-437. Romania. ERIC ED613253. Google Scholar

[71] Al-Jarf, R. (2009a). Effects of online collaborative activities on second language acquisition. 14th Annual TCC Worldwide Conference Online Conference. April 14-16. Google Scholar

[72] Al-Jarf, R. (2009b). *How to prepare English language tests*. College of Languages and Translation Symposium series. King Saud University. Riyadh, Saudi Arabia. December 26. Google Scholar

[73] Al-Jarf, R. (2008a). Acquisition of adjective-forming suffixes by EFL freshman students. TELLIS Conference, February 17-18. Islamic Azad University-Roudehen. https://doi.org/10.2139/ssrn.3842264. ERIC ED609956. Google Scholar

[74] Al-Jarf, R. (2008b). *Benchmarks for staffing translation departments in Saudi Arabia*. College of Languages and Translation 2nd Annual Meeting. King Saud University, Riyadh, Saudi Arabia. April 26-30. ERIC ED611785, Google Scholar

[75] Al-Jarf, R. (2008c). Is internet-based learning effective in EFL. In *13th TCC Online Conference. April* (pp. 16-18). Google Scholar

[76] Al-Jarf, R. (2008d). Phonological and orthographic problems in EFL college spelling. First Regional Conference on English Language Teaching and Literature (ELTL 1). Islamic Azad University-Roudehen. TELLIS Conference Proceedings. ERIC ED611115. Google Scholar

[77] Al-Jarf, R. (2008e). *Thesis evaluation challenges in Saudi Arabia as perceived by graduate students, advisors and examiners*. Conference on Peer Reviewing. Imam University, Riyadh, Saudi Arabia. Google Scholar

[78] Al-Jarf, R. (2007a). *A model for quality criteria for preparing secondary school students for university studies and life.* 14th annual Conference of the Saudi Educational and Psychological Association titled Quality in Education. P. 661-690. https://www.researchgate.net/publication/280796383. Google Scholar

[79] Al-Jarf, R. (2007b). *Assessing students' reading competencies: Setting global standards*. Asian EFL Journal Global Congress. Seoul, Korea. May 25-26. Google Scholar

[80] Al-Jarf, R. (2007c). Impact of blended learning on EFL college readers. IADIS International Conference on e-Learning, Lisbon. ERIC ED634492. Google Scholar

[81] Al-Jarf, R. (2007d). Online ESL Learning: effects on college levels & course types. College of Languages and Translation. https://www.researchgate.net/publication/267546695 . Google Scholar

[82] Al-Jarf, R. (2007). Processing of advertisements by EFL Arab college students. The Reading Matrix Journal, 7, 1, April, 132-140. Google Scholar

[83] Al-Jarf, R. (2007e). Teaching vocabulary to EFL college students online, Call-EJ Online 8 (2), 1-16. Google Scholar

[84] Al-Jarf, R. (2007f). *Testing reading for special purposes*. Conference on Assessing Language and (Inter-) cultural Competences in Higher Education Turku, Finland. August 30 - September 1. Google Scholar

[85] Al-Jarf, R. (2007g). *Testing research skills in EFL*. Conference on Assessing Language and (Inter-) cultural Competences in Higher Education Turku, Finland. August 30 - September 1. Google Scholar

[86] Al-Jarf, R. (2006a). Impact of online instruction on EFL students' cultural awareness. ERIC ED497400. Google Scholar

[87] Al-Jarf, R. (2006b). Plural acquisition by EFL freshman college students. GLOBE Conference entitled: "Communicating across age groups: Age, language and society. Warsaw, Poland. September 21-23. Google Scholar

[88] A-Jarf, R. (2006c). The effects of elearning on teaching English as a foreign language to Saudi college students. Mission of Education and Psychology Journal, 26, 215-242. Saudi Association for Education and Psychology, King Saudi University. Google Scholar

[89] Al-Jarf, R. (2005a). Task-based instruction for EFL struggling college writers. International Conference on Task-Based Language Teaching (TBLT 2005). Centre for Language and Migration, University of Leuven, Belgium. https://www.academia.edu/116296124/. Google Scholar

[90] Al-Jarf, R. (2005b). The effects of listening comprehension and decoding skills on spelling achievement of EFL freshman students. English language and literature Education. Journal of the English Language Teachers in Korea (ETAK), 11, 2. ERIC ED625524. Google Scholar

[91] Al-Jarf, R. (2005c). The effects of online grammar instruction on low proficiency EFL college students' achievement. Asian EFL Journal, 7, 4, 166-190. ERIC ED634096. Google Scholar

[92] Al-Jarf, R. (2005d). The relationship among spelling, listening and decoding skills in EFL freshman students. *English Language & Literature Teaching, Vol. 11, No. 2, 35-55.* Google Scholar

[93] Al-Jarf, R. (2004a). Differential effects of online instruction on a variety of EFL courses. The 3rd Annual Meeting of the Asia Association of Computer Assisted Language Learning (AsiaCALL), Penang, Malaysia, Nov 22-24. ERIC ED497936. Google Scholar

[94] Al-Jarf, R. (2004b). The effect of web-based learning on struggling ESL college writers. Foreign Language Annals, 37, 1, 46-56. DOI: 10.1111/j.1944-9720.2004.tb02172.x. Google Scholar

[95] Al-Jarf, R. (2003). *An analytical study of translation tests*. College of Languages and Translation Symposium Series, King Saud University. Riyadh, Saudi Arabia. December 1. Google Scholar

[96] Al-Jarf, R. (2002a). Effect of online learning on struggling ESL college writers. ERIC ED475920.

[97] Al-Jarf, R. (2002b). *Linguistic and measurement considerations in Translation tests.* 13th World Congress of the Association Internationale de Lingistique Appliquee (AILA). Singapore, December 16-21. Google Scholar www.researchgate.net/publication/350314137. Google Scholar

[98] Al-Jarf, R. (2002c). *Reflections on translation assessm*ent. American Association of Applied Linguistics (AAAL) Conference. Salt Lake City, Utah, April 6-9. www.researchgate.net/publication/350314093. Google Scholar

[99] Al-Jarf, R. (2001). *Issues in translation assessment. 5th* CTELT Annual Conference "Teaching, Learning and Assessment", Dubai, United Arab Emirates, May 9-10. www.researchgate.net/publication/350314112. Google Scholar

[100] Al-Jarf, R. (1998). *Evaluation of the EFL program at King Faisal schools: Grades 1-12.* https://www.researchgate.net/profile/R. -Al-Jarf/publication/280943034. Google Scholar

[101] Al-Jarf, R. (1995). *An Arabic word identification diagnostic test for the first three grades*. Center for Educational Research. College of Education. King Saud University. Google Scholar

[102] Al-Jarf, R. (1994). Analysis of Arabic first, second and third grade students' errors in word identification. *Journal of Contemporary Education; Cairo; 9(61)*, 88-147. Google Scholar

[103] Al-Jarf, R. (1989). *Criteria for evaluating graduate programs*. Proceedings of the Second Annual Symposium of the Graduate College. King Saud University, 103-126. ERIC ED638713. Google Scholar

[104] Ardian, T., Sudiana, I. & Putrayasa, I. (2025). Literature Review Study: Language Skills and Forms of Assessment. *Jurnal BELAINDIKA: Pembelajaran dan Inovasi Pendidikan*, *7*(2), 174-183.

[105] Bahi, H., Dendani, B., & Lounis, M. (2024). Automatic Pronunciation Assessment and Feedback for Arabic Learners: A Review. *International Journal of Asian Language Processing*, *34*(03n04), 2430001.

[106] Buchanan, K., et al. (2025). A systematic review of early writing assessment tools. *Early Childhood Education Journal*, *53*(6), 1939-1949.

[107] Carbonieri, J. & Lúcio, P. (2020). Vocabulary assessment in Brazilian children: a systematic review with three instruments. In *CoDAS,* 32, e20180245). Sociedade Brasileira de Fonoaudiologia.

[108] Chico, R. (2026). Effectiveness of Grammar-Integrated Authentic Assessment in Enhancing Language Proficiency among High School Students: A Meta-Analysis Review. *International Journal of Research and Innovation in Social Science*, *10*(1).

[109] Dang, H. & Habók, A. (2026). Digital reading comprehension assessment in English language education: a systematic review. *Cogent Education*, *13*(1), 2633011.

[110] Denman, D., et al. (2017). Psychometric properties of language assessments for children aged 4–12 years: A systematic review. *Frontiers in psychology*, *8*, 1515.

[111] Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies*, *29*(11), 14151-14203.

[112] Dixon, C., et al. (2023). Dynamic assessment as a predictor of reading development: a systematic review. *Reading and Writing*, *36*(3), 673-698.

[113] Djiwandono, P., & Ginting, D. (2025). Evaluating research reports on the qualities of tests of English language skills in Indonesian schools: A systematic review. *Language Education & Assessment*, *8*, 2237-2237.

[114] Dobinson, K. & Dockrell, J. (2021). Universal strategies for the improvement of expressive language skills in the primary classroom: A systematic review. *First Language*, *41*(5), 527-554.

[115] Dujardin, E., et al. (2021). Tools and teaching strategies for vocabulary assessment and instruction: A review. *Social Education Research*, 34-66.

[116] El Kheir, Y., et al. (2023). Automatic pronunciation assessment-a review. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8304-8324.

[117] Farhady, H. (2018). History of language testing and assessment. *The TESOL encyclopedia of English language teaching*, 1-7.

[118] Fulcher, G., Panahi, A., & Mohebbi, H. (2022). Glenn Fulcher's Thirty-Five Years of Contribution to Language Testing and Assessment: A Systematic Review. *Language Teaching Research Quarterly*, *29*, 20-56.

[119] Gillis, A., Morris, M., & Ridgway, P. (2015). Communication skills assessment in the final postgraduate years to established practice: a systematic review. *Postgraduate medical journal*, *91*(1071), 13-21.

[120] Honorato-Errázuriz, J., & Ramírez-Montoya, M. (2021, October). Randomized Evaluation of Reading Skills: An Opportunity for Systematic Literature Review. In *Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21)* (pp. 616-623).

[121] Hu, H., Gong, Q., & Said, N. (2025). Exploring a decade of research: A systematic review of computer-based English speaking tests. In *Forum for Linguisfic Studies* (Vol. 7, No. 4, pp. 788-803).

[122] Huawei, S., & Aryadoust, V. (2023). A systematic review of automated writing evaluation systems. *Education and Information Technologies*, *28*(1), 771-795.

[123] Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing*, *38*(2), 189-218.

[124] Marinho, A., et al. (2022). Public speaking assessment and self-assessment instruments: an integrative literature review. *Audiology-Communication Research*, *27*, e2539.

[125] McIntyre, L., et al. (2017). Receptive and expressive English language assessments used for young children: a scoping review protocol. *Systematic reviews*, *6*(1), 70.

[126] Muès, M., et al. (2024). Factors associated with receptive and expressive language in autistic children and siblings: A systematic review. *Autism & Developmental Language Impairments*, *9*, 23969415241253554.

[127] Nitami, K. & Santosa, M. (2024). A Systematic Literature Review: The Effectiveness of Vocabulary Size Tests on Students English Vocabulary and Writing Skills. *LETS: Journal of Linguistics and English Teaching Studies*, *6*(1), 1-13.

[128] Ntonti, P., et al. (2023). A systematic review of reading tests. *International Journal of Ophthalmology*, *16*(1), 121.

[129] Saptiany, S., & Prabowo, B. (2024). Speaking proficiency among English specific purpose students: a literature review on assessment and pedagogical approaches. *LITERACY: International Scientific Journals of Social, Education, Humanities*, *3*(1), 36-48.

[130] Sescleifer, A., Francoisse, C. & Lin, A. (2018). Systematic review: Online crowdsourcing to assess perceptual speech outcomes. *Journal of Surgical Research*, *232*, 351-364.

[131] Sprenger-Charolles, L., & Messaoud-Galusi, S. (2009). Review of research on reading acquisition and analyses of the main international reading assessment tools.

[132] Taufiqi, A., Sodiq, S., & Amri, M. (2025, September). Assessing Speaking Skills Through E-Assessment: A Systematic review for Advancing Language Evaluation in Indonesian Language Education. In *Proceeding of International Joint Conference on UNESA* (Vol. 3, No. 1, pp. 334-352).

[133] Trembath, D., Westerveld, M., & Shellshear, L. (2016). Assessing spoken language outcomes in children with ASD: A systematic review. *Current Developmental Disorders Reports*, *3*(1), 33-45.

[134] Usha, G. & Alex, J. (2023). Speech assessment tool methods for speech impaired children: a systematic literature review on the state-of-the-art in speech impairment analysis. *Multimedia tools and applications*, *82*(22), 35021-35058.

[135] Wind, S. & Peterson, M. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161-192.

[136] Wood, E., Biggs, K., & Molnar, M. (2024). Characteristics of dynamic assessments of word reading skills and their implications for validity: A systematic review and meta-analysis. *SAGE Open*, *14*(4), 21582440241300536.

[137] Xia, Y., Luo, Y., & Lu, X. (2024). Dynamic Assessment of Vocabulary: A Systematic Literature Review. *Language Teaching Research Quarterly*, *46*, 297-324.