British Journal of Teacher Education and Pedagogy
ISSN: 2755-1482
DOI: 10.32996/bjtep
Journal Homepage: www.al-kindipublisher.com/index.php/bjtep

BJTEP
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

An Integrative Systematic Review of Studies Across Diverse Educational Evaluation Domains

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

| ABSTRACT

This study conducted an integrative systematic review (ISR) synthesizing the author's research on educational evaluation published between 1989 and 2023. The corpus comprises sixteen studies spanning multiple evaluation domains, including thesis and graduate program evaluation, peer reviewing, curriculum and textbook evaluation, teacher performance appraisal, learning outcomes and course grade patterns, digital assessment tools, and admission benchmarks. The studies were organized into seven thematic clusters: MA/PhD theses and graduate program evaluation; peer review and publishing quality assurance; assessment, testing, and grade outcomes in language and translation education; instructor qualification evaluation; foreign language curriculum and textbook evaluation; digital assessment tools; and program admission policies and staffing benchmarks in language and translation education. Across clusters, the findings revealed a sustained scholarly effort to diagnose systemic weaknesses in educational evaluation across school, university, and professional contexts. The studies demonstrated inconsistent or unclear evaluation standards; variability in evaluator expertise; structural and administrative constraints that shape instructional and evaluative processes; and the significant influence of assessment policies on student learning, academic integrity, and institutional accountability. These themes recur across graduate research supervision, peer review systems, EFL curriculum design, textbook development, teacher performance evaluation, and course grade distribution, indicating that the challenges identified are structural rather than context specific. The ISR also highlights recurring patterns across the corpus as persistent misalignment between intended standards and actual practice, limited methodological rigor in some evaluation processes, and the far reaching consequences of institutional policies, particularly during periods of disruption such as the COVID 19 pandemic. The studies collectively demonstrate that weaknesses in evaluation practices often stem from systemic issues such as inadequate training, inconsistent regulations, insufficient oversight, and the absence of unified benchmarks across institutions. This ISR provides the first comprehensive synthesis of a 35 year research program on educational evaluation, revealing conceptual coherence and methodological evolution across diverse evaluation domains. By mapping patterns, challenges, and cross domain insights, this ISR offers evidence that can inform future evaluation frameworks, strengthen institutional policies, and guide reforms in curriculum design, teacher appraisal, graduate studies supervision, and assessment practices. The findings highlight the need for unified standards, enhanced evaluator preparation, and system level approaches to improving educational quality across sectors.

| KEYWORDS

Integrative Systematic review (ISR), Al-Jarf research program, educational evaluation, peer reviewing, teacher performance evaluation, curriculum and textbook evaluation, learning outcomes, digital assessment tools, admission benchmarks

| ARTICLE INFORMATION

ACCEPTED: 25 March 2026 **PUBLISHED:** 30 March 2026 **DOI:** 10.32996/bjtep.2026.5.3.5

1. Introduction

Evaluation[1] is the systematic process of collecting**,** analysing, and interpreting data to determine the effectiveness of teaching, curriculum, and student learning outcomes to make informed decisions for improvement. It moves beyond mere testing to determining if educational goals have been achieved, using formative and summative methods to improve instruction and measure achievement. According to Ralph W. Tyler[2], an American educator and psychologist, "*Evaluation is the process of determining the extent to which educational objectives are being realised.*"

Although assessment and evaluation are used interchangeably, assessment focuses on gathering information about student learning, whereas evaluation is the interpretation of that evidence to make a judgment on the effectiveness of the teaching and learning process. Evaluation therefore encompasses both formative and summative processes, using multiple sources of evidence to determine whether educational goals have been met.

Educational evaluation is applied in many domains as program and curriculum evaluation, textbook evaluation, teacher performance evaluation, student assessment and learning outcomes, program and institutional evaluation, thesis and dissertation evaluation, peer review and publishing quality assurance, admission and placement evaluation, digital and technology-enhanced assessment, teacher training and professional development evaluation.

Evaluation serves several purposes and functions: Improving instruction; measuring student achievement and determining if students have met learning goals; informing administrative decisions and understanding whether curriculum revisions and teacher training programs are needed; providing feedback to help students know their strengths, weaknesses and areas of improvement; enhancing curriculum effectiveness by matching learners' needs, and meeting educational standards; and ensuring that educational programs meet quality standards. Evaluation functions[3] cover relevance, efficiency, effectiveness, impacts, and sustainability, all aimed at assessing program performance and guiding improvements. Ultimately, evaluation ensures accountability and provides valuable information for decision-making.

Evaluation in Education goes through the following stages: Defining objectives, i.e., what students are expected to learn; selecting teaching/learning strategies; choosing the evaluation tools and methods as tests, rubrics, portfolios, observations etc; gathering information through tools and methods; analysing and interpreting data; making judgments and decisions whether the objectives are accomplished or not; providing and sharing Feedback for improvement if needed. Evaluation must be based on clearly defined learning objectives. Evaluation must cover skills, knowledge, attitudes, and behavior changes. Through evaluation, teachers diagnose problems and suggest solutions for improvement. Evaluation is cooperative.

Student evaluation in education takes different forms, each has a specific purpose and is used at different intervals in the teaching-learning process. It includes formative, summative, diagnostic, placement, achievement and aptitude evaluation and others. Teachers use tests and quizzes, performance-based tasks, observation, self-assessment and peer assessment, rubrics, oral examinations/interviews, case studies and group projects, surveys and questionnaires, performance-based assessments (projects, portfolios), teacher observations, classroom discussions, and formal tests. These methods provide a comprehensive view of student learning and program efficacy.

Given the centrality of evaluation in education, a review of the literature revealed a substantial body of research and numerous systematic reviews (SRs) addressing different dimensions of educational evaluation. One major group of SRs has examined evaluation models, curriculum evaluation, and broad program assessment. Examples of these evaluation SRs are: curriculum and program evaluation in medical education (Ullah et al., 2024); evaluation of the implementation of project-based learning in engineering programs (Ramírez de Dampierre et al., 2024); outcome-based assessment in the evaluation of education programs (Agir, N., et al., 2023); educational design and evaluation models of learning effectiveness in e-learning (Spatıotı et al., 2023); postgraduate theses on curriculum evaluation (Tas & Duman, 2021); the efficacy of three program-evaluation models (Iqbal et al., 2021); recent trends in educational program and curriculum evaluation models (Nouraey et al., 2020); evaluations of reproductive health programs in humanitarian settings (Casey, 2015); how evaluation and audit is implemented in educational organizations (Farahsa & Tabrizi, 2015); and evaluation of educational administration (Parylo, 2012); qualitative evaluation methods in ethics education (Watts, et al., 2017).

[1] https://www.unesco.org/en/query-list/e/evaluation-education
[2] https://www.21kschool.com/sa/blog/what-is-evaluation-in-education/
[3] https://www.gov.scot/publications/5-step-approach-evaluation-designing-evaluating-behaviour-change-interventions/pages/4/

A second group of SRs focused on evaluation in health-related, nursing, and medical education contexts. These include evaluation SRs of the validity and reliability of tools in online OSCE exams (Felthun et al., 2021); baccalaureate nursing programs (Al-Alawi & Alexander, 2020); students' mental and social health promotion educational programs (Baghian et al., 2019); dissertations in health service psychology programs (Vidair et al., 2019); technology-enhanced learning programs for health care professionals (Nicoll et al., 2018); formal continuing medical education (Tian et al., 2007); evaluation of online graduate nursing education (Horne & Sandmann, 2012).

A third group of SRs examined evaluation in EFL/ESL programs, teacher development, and assessment literacy. These include EFL online assessment in higher education (Wannas & AbdelMohsen, 2025); EFL teachers' language assessment literacy (Puspawati et al., 2024); authentic assessment in k-12 ESL/EFL education (Osman, 2023); EFL teachers' training needs for professional development programs (Alzahrani & Nor, 2021); program evaluation studies in EFL in Turkey (İpek, 2022); and status of EFL teachers' professional development in Turkey (Hos & Topal, 2013);

Another group of SRs focused on the evaluation of theses, dissertations, and academic research output, such as the quality, characteristics, or evaluation of graduate research as in research on dissertations in health service psychology programs (Vidair et al., 2019); evaluability assessment thesis and dissertation studies in graduate professional degree programs (Walser & Trevisan, 2016); tourism research knowledge and cross-cultural evaluation of doctoral theses (Oliveira et al., 2015); and postgraduate research at the university of Zambia: a review of dissertations for the master of medicine programme (Ahmed et al., 2010); and Master's medical research at the university of Zambia (Ahmed et al., 2010).

Further SRs analyzed peer-review processes, reviewer tasks, and quality assurance in scholarly publishing as quality in peer review reports (Sizo et al., 2025); feedback practices in journal peer-review (Chong & Lin, 2024); online peer-review and assessment systems (Babik et al., 2024); online training in manuscript peer review (Willis et al., 2022); peer review of searches for studies for health technology assessments (Lefebvre & Duffy, 2021); the roles and tasks of peer reviewers in biomedical journals (Glonti et al., 2019); and peer review systems (Lasker, 2018).

A final group of SRs addressed online assessment, proctoring technologies, cheating, and exam integrity. These include SRs as academic integrity through automated online exam proctoring (Malhotra & Chhabra, 2026); open book exams in higher education (Shunko, 2025); web-based examinations in higher education institutions in Sub-Saharan Africa (Bervell et al., 2025); students' acceptance of and preferences regarding online exams (Topuz & Kinshuk, 2024); the efficacy of online proctoring in online examinations (Kuleva & Miladinov, 2024); online exams in physics and maths (Braun, 2024); examinees' identity authentication in online distant exams (Saleh et al., 2023); automated online exam proctoring approaches (Fatima et al., 2022); deep learning-based online exam proctoring systems (Abbas & Hameed, 2022); online exams in Turkey (Albayrak, 2022); machine learning models for online learning and examination systems (Kaddoura et al., 2022); online exams in the Covid-19 pandemic (Dayananda et al., 2021); cheating in online exams (Öncül, 2021; Năznean, 2021); online exams solutions in e-learning (Muzaffar et al., 2021); and take-home exams in higher education (Bengtsson, 2019); students' acceptance of and preferences regarding online exams (Topuz & Kinshuk, 2024).

Despite the existence of numerous systematic reviews on isolated aspects of educational evaluation—such as EFL/ESL assessment, curriculum evaluation, online assessment technologies, program evaluation in medical education, dissertation quality, and peer-review systems—the literature remains highly fragmented. Each review focuses on a single domain, population, or context, leaving no integrative synthesis that maps how evaluation practices, methods, and outcomes intersect across the broader field. As a result, there is no comprehensive understanding of how educational evaluation has evolved over time, where evidence converges or diverges, or what gaps persist across domains such as teacher evaluation, curriculum and textbook evaluation, learning-outcome assessment, program evaluation, grade distribution, and digital assessment tools. To address this gap, the present study aims to conduct an integrative systematic review (ISR) that synthesizes the author's research studies on educational evaluation published between 1989 and 2023. The studies cover thesis and graduate-studies evaluation, peer reviewing, curriculum and textbook evaluation, teacher performance appraisal, learning outcomes, learning outcomes and course grade patterns, digital assessment tools, and admission benchmarks.

This study is significant because it brings together, for the first time, a coherent body of research produced over 35 years by a single scholar which have never been examined as an integrated whole. By synthesizing studies that span multiple evaluation domains—curriculum and textbook evaluation, teacher performance appraisal, thesis and graduate-studies evaluation, peer-review and academic quality assurance, assessment practices, learning outcomes, and admission benchmarks—the review reveals the

conceptual continuity, methodological development, and cumulative contributions of the author's work. It transforms a set of independent studies into a unified, cross-disciplinary map of educational evaluation in Arab and international contexts, highlighting recurring patterns, systemic challenges, and areas of convergence across domains. This integrative perspective provides researchers, policymakers, and practitioners with a consolidated evidence base that can inform future evaluation frameworks, institutional reforms, and strategic decision-making.

Finally, this ISR is significant because it is part of a broader series of SR/MA projects by the author, that has so far included the following SRs/MAs of studies on *an integrative review of studies on teaching English for art education purposes to ph.d. students* (Al-Jarf, 2026a); *a researcher's contributions to EFL reading instruction: themes, methods, and pedagogical insights* (Al-Jarf, 2026b); *translation error studies (2000–2025): the case of students' errors in English–Arabic and Arabic–English translation* (Al-Jarf, 2026c); *mobile apps for developing multiple language skills in EFL* (Al-Jarf, 2026d); *adult reading practices, interests, habits and challenges* (Al-Jarf, 2026e); *pronunciation instruction and practice in l2 (2005–2025)* (Al-Jarf, 2026f); *teaching reading in Arabic to grades 1–12: textbooks, skills, and learning outcomes* (Al-Jarf, 2026g); *Arabic–English transliteration of personal names and public signages* (Al-Jarf, 2026h); *children's language acquisition and development in Saudi arabia* (Al-Jarf, 2026i); *classroom practices, writing enhancement and creativity among EFL struggling students* (Al-Jarf, 2026j); *collaborative learning and teaching in digital environments* (Al-Jarf, 2026k); *effectiveness of mind-mapping on multiple English language skills in the Saudi context* (Al-Jarf, 2026l); *inadequate staffing and large class sizes in Saudi EFL and translation programs* (Al-Jarf, 2026m); *innovative word formation and pluralization processes in Arabic* (Al-Jarf, 2026n); *2024–2025 studies on ai Arabic translation, linguistics and pedagogy* (Al-Jarf, 2026o); *ESP innovation* (Al-Jarf, 2026p).

2. Methodology

2.1 Study Corpus

The study corpus consists of 16 studies by the author published between 1989 and 2023 in a range of international journals conference proceedings, book chapters and reports. The studies were included in the current ISR if they met the following criteria: (i) The study must be authored or co-authored by Reima Al-Jarf. (ii) Participants must be graduate and undergraduate students, EFL teachers, translation instructors, and evaluation specialists. (iii) The study must address an educational evaluation domain, as thesis, graduate and EFL programs, L2 curriculum and textbooks, translation, language, linguistic, education, computer science courses, assessment tools, and admission benchmarks. (iv) The study must be published between 1989 and 2023, reflecting the full span of the author's research program. (v) The publications include peer-reviewed journal articles, conference papers, book chapters and reports. (vi) Studies published in English or Arabic were included. The full text must be accessible for analysis.

Cluster 1: MA/PhD theses and Graduate Program Evaluation

Studies focusing on the quality of MA/PhD theses, evaluation criteria, supervision issues, and characteristics of graduate research, evaluation standards, thesis quality, supervision challenges, institutional differences, and improvement strategies. These studies include:

- *MA and Ph.D. thesis evaluation at Saudi universities: problems and solutions (Al-Jarf, 2022c)*
- *Characteristics of Ph.D. Dissertations of Saudi students who graduated from American universities between 1969–1985 (Al-Jarf, 1991)*
- *Criteria for evaluating graduate programs (Al-Jarf, 1989)*

Cluster 2: Peer Review and Publishing Quality Assurance

Studies addressing the challenges faced by peer reviewers in local and international academic contexts, reviewer workload, quality criteria, institutional pressures, and differences between local and international review systems. These studies include:

- *Challenges faced by Arab peer-reviewers (Al-Jarf, 2023a)*
- *Challenges faced by peer-reviewers for local and international academic institutions (Al-Jarf, 2019)*

Cluster 3: Assessment, Testing, and Grade Outcomes in Language and Translation Education
A) Grade Outcomes and Grade Inflation

These studies focus on temporal comparisons, pandemic effects, disciplinary differences, and systemic grading patterns. They include:

- *grade inflation at Saudi universities before, during and after the pandemic: a comparative study (Al-Jarf, 2022b)*
- *grade inflation in language and translation courses at Saudi schools and universities (Al-Jarf, 2022a)*

B) Assessment and Testing Practices

These studies focus on test validity, assessment design, e-assessment challenges, and pedagogical implications. They include:

- *critical analysis of translation tests in 18 specialized translation courses: shortcomings and recommendations (Al-Jarf, 2021a)*
- *online exams in language, linguistics and translation courses during the pandemic in Saudi Arabia (Al-Jarf, 2022d)*

Cluster 4: Evaluation Instructor Qualifications
Studies focusing on evaluating individual instructors' performance, teaching quality, teaching effectiveness, digital evaluation tools and rubric-based assessment. They include:

- *Assessing EFL college instructors' performance with digital rubrics (Al-Jarf, 2015a)*
- *Role of instructor qualifications, assessment and pedagogical practices in EFL students' grammar and writing proficiency (Al-Jarf, 2022).*

Cluster 5: Evaluation of Foreign Language Curriculum and Textbooks
Studies evaluating **curricula, academic programs, and institutional effectiveness**, program outcomes, and standards for academic program quality. These include:

- *Evaluation of Russian Arabic language teaching textbooks in the light of CEFR criteria (Al-Jarf & Mingazova, 2020a).*
- *How much material do EFL college instructors cover in reading courses (Al-Jarf, 2021c)*
- *Evaluation of the EFL program at king Faisal schools: grades 1–12 (Al-Jarf, 1998*

Cluster 6: Digital Assessment Tools
- *Creating and sharing iRubrics using RCampus (Al-Jarf (2010)*

Cluster 7: Program Admission Policies and Staffing Benchmarks in Language and Translation Education
- *Benchmarks for staffing translation departments in Saudi Arabia (Al-Jarf, 2008).*

1.2 *Eligibility (Inclusion & Exclusion) Criteria*

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

- **Duplicate studies of previously published work without adding new data or analysis**. These studies overlap conceptually and methodologically with other included works and, therefore, were excluded to avoid redundancy. Examples are: *MA and Ph.D. thesis evaluation problems and proposed solutions (Al-Jarf, 2008a); an analytical study of translation tests (Al-Jarf, 2003); linguistic and measurement considerations in translation tests (Al-Jarf, 2002a); reflections on translation assessment (Al-Jarf, 2002b); and issues in translation assessment (Al-Jarf, 2001); thesis evaluation challenges in Saudi Arabia as perceived by graduate students, advisors and examiners (Al-Jarf, 2008b)*

- **Studies on a single course or single skill assessment.** These studies focus on assessment within a single course, single skill, or single instructional context, which does not align with the broader institutional/program-level evaluation scope of the current systematic review. Examples are: *Testing multiple vocabulary associations for effective long-term learning (Al-Jarf, 2023c); standardized test preparation with mobile flashcard apps (Al-Jarf, 2021d); testing reading for specific purposes in an art education course for graduate students (Al-Jarf, 2021e; Al-Jarf, 2021f); EFL female college students and*

instructors' preferred method of speaking assessment (Al-Jarf, 2021b); how EFL college instructors can create and use grammar irubrics (Al-Jarf, 2020b); empowering EFL teachers and students with grammar irubrics (Al-Jarf, 2011c); creating and sharing vocabulary irubrics (Al-Jarf, 2012); *what teachers should know about reading tests (Al-Jarf, 2017a); what teachers should know about vocabulary tests for EFL freshman students (Al-Jarf, 2015d); Issues in assessing the speaking skill in EFL (Al-Jarf, 2015b); test preparation with mobile apps (Al-Jarf, 2014); Assessing graduate students' research skills in EFL (Al-Jarf, 2013); developing and testing reading skills through art texts (Al-Jarf, 2011; Al-Jarf, 2011b); how to prepare English language tests (Al-Jarf, 2009); assessing students' reading competencies: setting global standards (Al-Jarf, 2007b); testing reading for special purposes (Al-Jarf, 2007c); testing research skills in EFL (Al-Jarf, 2007d); analysis of Arabic first, second and third grade students' errors in word identification (Al-Jarf, 1994).*

- **Evaluation Checklists, iRubrics and standardized tests** such as *Textbook evaluation checklist (Al-Jarf, 2015c); An Arabic word identification diagnostic test for the first three grades (Al-Jarf, 1995); Creating and sharing writing iRubrics (Al-Jarf, 2011a)*

- **Studies where evaluation is a partial component** as *Preparing high school students for the university and life after graduation (Al-Jarf, 2023b);* a *model for quality criteria for preparing secondary school students for university studies and life (Al-Jarf, 2007a).*

2.1 *Corpus Characteristics*

The final corpus consisted of 16 studies authored by Reima Al-Jarf between 1989 and 2023. Because the dataset represents a closed, author-bounded research program, it is both comprehensive and internally coherent, reflecting the author's sustained scholarly trajectory in the evaluation of graduate theses and graduate programs, peer-reviewing practices, L2 and translation programs, curriculum and textbooks, learning outcomes, grade inflation, teacher performance, admission benchmarks, and digital assessment tools. Although the studies address diverse topics, they share a consistent analytical orientation and employ comparable data-collection approaches. Across the corpus, data were primarily descriptive and derived from three sources: questionnaire surveys and interviews with undergraduate and graduate students, thesis supervisors, journal peer reviewers, and ESL and translation teachers, rating scales, performance rubrics and students' final course grades, and textbook content analysis. Together, these methods generated a coherent body of evidence on educational evaluation practices and outcomes across Saudi Arabia and international contexts. For synthesis purposes, the 16 studies were grouped into 7 thematic clusters, each representing a distinct dimension of the research program. Collectively, these clusters offer an integrated overview of the author's contributions to educational evaluation. Overall, the corpus reflects a longitudinal, methodologically consistent body of work that traces the evolution of educational practices, outcomes, and evaluation domains over time.

2.2 *Information Sources*

The information sources for this ISR 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 educational evaluation 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, ResearchGate, Semantic Scholar, Academia.edu, SSRN, ERIC, EBSCO, ProQuest, and institutional repositories. Collectively, these platforms provide full coverage of the author's publications across journals, book chapters, 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.

2.3 *Data Extraction and Synthesis*

Because the corpus represents a single author's long-term research program, the methodological framing and analytical categories were highly consistent across all included studies. This internal coherence minimized coding discrepancies and enabled a unified synthesis of findings spanning 35 years of scholarly work.

Data extraction and synthesis followed an integrated, multi-stage procedure tailored to the descriptive and heterogeneous nature of the studies. For each study, information was extracted directly from the full text, including publication year; research domain (e.g., thesis supervision, Saudi PhD students' theses in education, graduate program evaluation, EFL curriculum evaluation for grades 1–12, Interactions I & II and Mosaic I & II, AFL textbook evaluation, translation course grades, teacher performance appraisal,

admission benchmarks, and content analysis); participant characteristics (undergraduate and graduate students, EFL teachers, translation instructors, thesis supervisors, peer reviewers, program evaluators); methodological approach (qualitative and quantitative analyses); data sources (frequency counts and percentages); and key findings related to EFL learning, translation course performance, and learning outcomes in language, translation, education, and computer science. These categories were selected to support thematic synthesis and cluster-level comparison rather than effect-size calculation, as the corpus consists predominantly of qualitative, descriptive studies in educational evaluation. All extracted information was entered into a structured matrix to ensure consistency and enable systematic comparison. Manual coding was used to preserve conceptual accuracy and classify each study according to the evaluation dimension it addressed.

Data synthesis proceeded in three stages. First, all studies were grouped into seven thematic clusters based on their primary focus: (1) MA/PhD theses and graduate program evaluation; (2) peer review and publishing quality assurance; (3) assessment, testing, and grade outcomes in language and translation education; (4) evaluation of instructor qualifications; (5) evaluation of foreign-language curricula and textbooks; (6) digital assessment tools; and (7) program admission policies and staffing benchmarks in language and translation education (see Section 2.1). This clustering enabled synthesis within conceptually coherent domains while preserving the distinct contributions of each study. Second, studies within each cluster were compared in terms of evaluation domains, data sources, recurring patterns, and pedagogical implications. Third, findings were synthesized across clusters to identify broader trends in educational evaluation.

Because the corpus reflects a single author's sustained research trajectory, the methodological consistency across studies strengthened the reliability of the synthesis and facilitated the integration of findings across a 35-year span.

2.4 *PRISMA Flow Description*

Because this ISR is based on a closed, predefined corpus of eleven studies published by the same author between 1989 and 2023, the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow reflects a streamlined identification and screening process. All publications within this time frame were retrieved from the academic platforms listed in Section 2.4 and manually screened for relevance. Each record was assessed against the eligibility criteria, and studies were excluded if they were duplicates or if they only partially addressed evaluation, focused on a single skill or single course, or represented isolated textbook evaluations or checklist-based assessments.

Following full-text screening, only studies that directly examined thesis evaluation, peer reviewing, teacher performance appraisal, curriculum and textbook evaluation, learning outcomes, digital assessment tools, and admission benchmarks were retained. The final set of studies was then organized into seven thematic clusters for synthesis. Accordingly, the PRISMA flow documents the progression from the initial identification of all publications within the author-bounded corpus, through screening and eligibility assessment, to the final inclusion of studies that substantively contribute to the evaluation domains central to this review.

3. Results

3.1 Overview

This ISR synthesizes findings from sixteen studies examining multiple domains of educational evaluation. The analysis is organized around seven 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 investigate key evaluation domains, including thesis evaluation, peer reviewing, teacher performance appraisal, curriculum and textbook evaluation, learning outcomes, digital assessment tools, and admission benchmarks.

This overview presents the major trends emerging from the corpus and illustrates how each thematic cluster contributes to a deeper understanding of educational evaluation. Collectively, the studies document the nature of evaluation practices, the tools and standards used, and the systemic issues that shape assessment across language, translation, and educational contexts.

3.2 Study Characteristics

The corpus consisted of 16 unique studies distributed over 7 thematic clusters. Findings of each cluster are presented below.

Cluster 1: Evaluation of MA/PhD Theses and Graduate Program Evaluation

MA and Ph.D. thesis evaluation at Saudi universities: problems and solutions (Al-Jarf, 2022c)

This study examines the challenges faced by graduate students, supervisors, co-advisors, and examiners during the evaluation and review of MA and PhD theses at three Saudi universities. Drawing on survey data from 5 MA and 4 PhD students, 10 supervisors, and 10 internal and external examiners, the findings reveal a wide spectrum of structural, academic, and ethical issues that undermine the quality and

credibility of postgraduate research. Student-related problems include weak research and electronic search skills, limited academic and linguistic competence, persistent methodological and writing errors, missed deadlines, insufficient effort, and resistance to revision. Some students attempt to influence examiners through personal relationships or resort to academic dishonesty, including purchasing ready-made studies or copying previous research.

Supervisor-related issues arise when supervisors lack specialization in the thesis topic, delay reviewing student work, provide vague or inconsistent feedback, or fail to commit to supervision hours. Frequent changes in supervisors and unclear expectations further complicate the process, and in some cases, supervisors end up writing parts of the thesis themselves. At the departmental and administrative levels, problems include assigning non-specialists to review proposals, rejecting topics without justification, inconsistent application of regulations, and delays in approving proposals or forming committees. Bias, favoritism, and personal conflicts sometimes influence decisions, compromising fairness and academic integrity.

Examiner-related challenges emerge during thesis defenses, where students may reject feedback or display weak methodological and linguistic skills. Supervisors may select friendly or non-specialist examiners to avoid criticism, and personal relationships can shape the tone and outcome of the defense. Departments may override external examiners' reports, form new committees to dilute critical evaluations, or approve weak theses under pressure. A major systemic issue is the absence of clear, unified, and objective evaluation standards; existing criteria are vague, inconsistently applied, and often influenced by personal preferences or administrative pressures. Additional concerns include the proliferation of commercial research services, social pressures on supervisors and examiners, students seeking external help without following feedback, and various forms of plagiarism. Collectively, the findings highlight deep-rooted academic, administrative, and ethical problems that compromise the rigor, fairness, and credibility of thesis supervision and evaluation in Saudi graduate programs.

Characteristics of Ph.D. Dissertations of Saudi students who graduated from American universities between 1969–1985 (Al-Jarf, 1991)

This study provides a descriptive analysis of doctoral dissertations in education completed by Saudi students at American universities between 1969 and 1985. The analysis examined multiple characteristics, including the number of dissertations produced, degree types, awarding institutions, dissertation length, authors' gender, research topics, educational levels investigated, research problems, methodologies, samples, statistical procedures, and major findings. Results showed that only five women earned doctoral degrees during this period. Most dissertations were completed at Michigan State University, the University of Northern Colorado, and Indiana University. The typical dissertation averaged 201 pages.

More than half of the dissertations focused on four areas: higher education, curriculum and instruction, educational administration, and science education. Approximately 35.71% examined issues at the university level, and 92.47% addressed traditional or previously explored topics. Only 21.43% of dissertation titles demonstrated strong writing quality. Methodologically, 68.62% of the dissertations were descriptive, 23.85% correlational, 7.53% causal, and only 4.6% experimental. The questionnaire was the most frequently used data-collection tool, appearing in 66.49% of the dissertations. The study concludes by recommending periodic analyses of doctoral dissertations, approximately every five years, to monitor evolving research trends and methodological patterns in Saudi graduate scholarship.

Criteria for evaluating graduate programs (Al-Jarf, 1989)

The article provides a conceptual analytical framework for evaluating graduate programs by synthesizing criteria from previous evaluation studies. It identifies eight major domains that should guide program evaluation: (i) Program Objectives, (ii) Need for the Program, (iii) Program Cost, (iv) Financial Capabilities, (v) Faculty Members, (vi) Academic Courses, (vii) Academic Atmosphere, and (viii) Students' Academic Level. Each domain is translated into a detailed checklist using five-point rating scales to assess how well a program meets each criterion. The study emphasizes the lack of unified national standards and argues that structured, multidimensional criteria are essential for ensuring quality, supporting decision-making, and improving the effectiveness of graduate programs.

Cluster 2: Peer Review and Publishing Quality Assurance

4+5) Challenges faced by Arab peer-reviewers (Al-Jarf, 2023a) and Challenges faced by peer-reviewers for local and international academic institutions (Al-Jarf, 2019)

Together, the two studies provide a comprehensive picture of the challenges faced by Arab peer reviewers in local, regional, and international academic contexts. Drawing on survey responses from 40 reviewers across several Arab countries (2023a) and combined telephone interviews and questionnaires with local and international reviewers (2019), the findings reveal a consistent

set of systemic problems that hinder the quality and efficiency of the peer-review process. Reviewers frequently receive manuscripts with significant linguistic, methodological, and structural weaknesses, which increases the burden of review and prolongs decision-making. Evaluation standards vary widely across journals and institutions, and many journals lack clear reviewing policies, transparent acceptance criteria, or standardized evaluation forms. The studies also highlight substantial variation in reviewer expertise, experience, and rigor; many reviewers are lenient or insufficiently specialized, while others face pressure from editors, authors, or institutional expectations.

Across both studies, reviewers report long turnaround times, inadequate follow-up from journals, and unclear or inconsistent editorial decisions regarding acceptance and rejection. Some journals struggle with a limited pool of qualified reviewers, while others fail to communicate their scope, specialization areas, or acceptance rates. Reviewers also describe external pressures—professional, social, or institutional—that can influence the objectivity of the review process. Collectively, the two studies show that peer-review challenges in the Arab academic sphere are structural and systemic, spanning manuscript quality, reviewer preparedness, journal policies, and editorial practices. Both studies conclude with recommendations for improving transparency, standardizing evaluation criteria, strengthening reviewer training, and enhancing journal management practices.

Cluster 3: Assessment, Testing, and Learning Outcomes in Language and Translation Coruses

A) Grade Outcomes and Grade Inflation

6) Grade inflation at Saudi universities before, during and after the pandemic: a comparative study (Al-Jarf, 2022b)

The study investigated grade inflation in 127 language, linguistics, translation, education, and computer courses taught over 8 semesters at some Saudi universities before, during and after the Pandemic. It was found that between 20% & 65% of the students chose a pass/no-grade results, the rest of the students mostly earned A & B grades in Spring 2020 when instruction and assessment were held online compared to students' grade in Fall 2018, Spring 2018, Fall 2019, and after the Pandemic (in Fall 2020, Spring 2021, Fall 2021, Spring 2022). Grade inflation was the highest in computer courses, followed by education courses and was the least in language, linguistics, and translation courses. Grade inflation in Spring 2020 was due to the adjustments mandated by universities to alleviate students' anxiety caused by the sudden shift to online teaching and assessment. Universities allocated 20% of the course marks to final exams, gave alternatives to a written final and were lenient in grading. Students had the option to drop the course or have a pass with no-grade result. In Fall 2020, classes were still held online but exams were held on campus. Starting Fall 2021, both instruction and exams were held on campus. Mark distribution and exam requirements went back to normal as before the Pandemic. However, grade inflation continued in many courses even in Spring 2022. The study gives recommendations for maintaining exam reliability, validity, and fairness in emergency and normal situations to achieve the desired learning outcomes.

7) Grade inflation in language and translation courses at Saudi schools and universities (Al-Jarf, 2022a)

This study investigated the status of grade inflation in language and translation courses in Saudi Arabia. Analysis of the pass rates and percentages of students who obtained Grades A+, A, B+ and B in 70 English language skills and translation college courses, in addition to the English course scores of students in grades 1 to 11 at a private school, showed evidence of grade inflation at the school and college levels as revealed by the high pass rates and high percentages of students obtaining Grades A+, A, B+and B in most courses. Responses to a questionnaire-survey showed several factors contributing to grade inflation such as administrators' tendency to raise students' marks, exercising pressure over instructors, correlating high quality with high pass rates, worrying about losing theirjob and parents' complaints and about being if students fail, giving easy questions, covering a small portion of the course/textbook material, lenient grading, allocating 60% of the course marks to attendance, assignments, quizzes, and class work, covering some course topics. Prior to exams, students enrolled in General English courses are given practice tests with similar questions to the final exam (in form and content). Hence, students know what to expect on the final exam. The study gives some recommendations for combating gradeinflation at Saudi schools and universities.

B) Assessment and Testing Practices

8) Critical analysis of translation tests in 18 specialized translation courses (Al-Jarf, 2021a)

The study describes and evaluates the current assessment practices prevalent in 18 translation courses offered at the College of Languages and Translation (COLT). Analysis of the # of English and Arabic source texts included, readability and difficulty level of texts included, # of exams with a terminology subtest, English and Arabic text length in words, and reliability, validity and discriminating power of final exams showed that 50% of the exams included one English text, 32% included 2 texts and 18% included 3 texts. 59% included one Arabic text, 9% included two texts and 5% included 3 texts; 56% included a vocabulary subtest.

41% did not have any Arabic texts. The English median text length was 181 words and the range 66-430 words. The median Arabic text length was 97 words and range 26-180. The typical Flesch Reading Ease of English texts was 40 and the typical Flesch-Kincaid Grade level score was 11. No significant differences existed among the different college levels or subject areas in text length or text difficulty level. The exams lack validity, reliability, and discriminating power. A model for more valid, reliable, and discriminating translation exams is given, with students' views of it.

9) Online exams in language, linguistics and translation courses during the pandemic in Saudi arabia (Al-Jarf, 2022d)

This study focused on online exams in language, linguistics, and translation courses in the first two semesters of the Pandemic (Spring 2020 and Fall 2020). Analysis of faculty surveys and students' comments on Twitter showed that the main concern of 91% of the students was final exams, passing courses with high grades, and the negative effect of online exams on their GPA, whether they would do well on Blackboard exams. Some cheated on online exams as their cameras were turned off. Numerous adjustments were mandated by university administrations to alleviate students' anxiety such as allocating 20% mark to the final exam, allowing more exam time, giving projects, open-book exams, term papers, reports, assignments or giving a presentation instead of the final. Some instructors gave easy questions, were lenient in grading, no essay, just objective questions, giving students the option to drop the course, to choose a letter grade, pass/fail, i.e., no grade, or to have a course mark included in their GPA. The study reports challenges of online exams during the Pandemic, design and delivery of online exams, assessment forms and choices, grade inflation issues, lessons learned and some recommendations.

Cluster 4: Instructor Evaluation

10) Assessing EFL college instructors' performance with digital rubrics (Al-Jarf, 2015a)

Many program coordinators and college administrators at the College of Languages and Translation are unable to accurately and dispassionately assess instructors' performance. There are discrepancies and biases in their evaluation reports. As a result, instructors are unhappy and feel underrated. Beginning and new instructors are not informed of what is expected of them. For those reasons, this article proposes the use of digital rubrics to ensure the reliability of teacher performance assessments. A rubric is a scoring guide that consists of specific pre-established performance criteria, used for evaluating students' and teachers' performance. In this article digital rubrics have been created for evaluating EFL college teachers' linguistic and professional competencies using the iRubric building tool of RCampus LMS. The author redesigned the existing TPA forms into two comprehensive digital rubrics: one for administrators and one for students, each organized into major competency categories.

The administrator rubric consolidates overlapping items from previous forms into 26 criteria grouped under six categories: linguistic competence, teaching performance, professional achievements, academic services, relationships with others, and personal qualities. Each criterion is accompanied by four performance levels (poor, fair, very good, excellent), each defined with detailed behavioral descriptors and point values. The student rubric retains the original 27 items but reorganizes them into six categories—teaching skills, assessment, feedback, punctuality, relationship with students, and overall opinion—again with clearly defined performance levels. Together, these rubrics offer a structured, evidence-based framework that supports fairer evaluations, clearer expectations, and more meaningful professional development for EFL instructors.

Finally, the article argues that digital rubrics enhance the validity and reliability of teacher assessment by setting explicit standards, clarifying expectations, and providing consistent scoring procedures. They also support program improvement by helping administrators diagnose weaknesses, plan targeted professional development, and align teaching practices with institutional goals. In this way, digital rubrics serve not only as evaluation tools but also as mechanisms for improving instructional quality and fostering a culture of transparency and accountability.

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

Three groups of EFL freshman students were concurrently enrolled in a grammar and a writing course. One group was taught the grammar and writing courses by the same instructor; the other two groups were taught grammar and writing by two different instructors using the same textbook but different instructional and assessment techniques. Comparisons of the grammar and writing post-tests scores showed significant differences between the three groups in the writing and grammar. There were strong correlations between the grammar and writing post-test scores. Performance of the Group that received a combination of writing and grammar instruction by the same instructor (Instructor A) was the highest. The relationship between grammar and writing instruction seems to be reciprocal: writing instruction affects grammatical competence and grammatical knowledge affects writing skill development. Better achievements were made when both courses were taught by the same instructor, as she can make the right connections between what is taught in both courses, and which specific structures and skills should be emphasized. The

instructors' qualifications, pedagogical system, educational and professional experience, the integration of online instruction, the type of error correction and instant feedback given to the students and the formative assessment technique used were significantly more effective than writing/grammar instruction that depended on the textbook alone. These variables proved to be important for enhancing the grammatical knowledge and writing quality of unskilled, low ability EFL students and resulted in a significant improvement in their grammar and writing scores.

Cluster 5: Evaluation of L2 Curriculum and Textbooks

12) Evaluation of Russian Arabic language teaching textbooks in the light of CEFR criteria (Al-Jarf & Mingazova, 2020a).

This study evaluated the textbooks titled *"Arabic for Non-native Speaking Children" I* & *II* used in teaching Arabic to elementary school children in Tatarstan in the light of the *Common European Framework of Reference* (CEFR) criteria. Results showed that the textbooks do not meet the CEFR criteria, as they focus on the reading and writing skills, not oral skills and communication. They also adopt a grammar-translation approach, not a communicative, functional approach to develop real-life language use. They present letters, vocabulary, and grammatical rules in isolation, with explanations in Russian. They do not support oral interaction, spontaneous communication, or the development of basic conversational abilities. They rely heavily on alphabet drills, decoding, penmanship, and isolated vocabulary lists. Lessons introduce words with pictures but rarely provide meaningful contexts, dialogues, or tasks. Reading passages are almost entirely absent from Level I and appear only at the end of Level II. Exercises are repetitive and mechanical, focusing on recognition and copying rather than comprehension, production, or communicative practice. They textbooks follow a structural, form-focused sequence. Grammar topics are presented explicitly and explained in Russian. Vocabulary selection is tied to the alphabet sequence or grammatical categories rather than communicative needs.

13) How much material do EFL college instructors cover in reading courses (Al-Jarf, 2021c)

EFL students at the College of Languages and Translation take 4 reading courses in the first four semesters of the translation program. The textbooks used are Interactions 1 & 2 and Mosaic 1 & 2. The study examines the amount of reading texts, reading exercises, and reading subskills covered by instructors in the Reading courses. Subjects were 24 instructors (6 instructors per course). Since students usually mark texts, do exercises, and take notes on their textbooks, three reading textbooks per instructor were randomly collected from students enrolled in the four Reading courses. Each book was examined page by page. The typical instructor taught 50% of the reading texts in Interactions 1 and Interactions 2; one third of the reading texts in Mosaic 1; and one fifth of the reading texts in Mosaic 2; 65% of the reading subskills and exercises in Interactions 1; half in Interactions 2; one third in Mosaic 1; and one fourth in Mosaic 2.

14) Evaluation of the EFL program at king Faisal schools: grades 1–12 (Al-Jarf, 1998)

This study is a comprehensive evaluative analysis of the English as a Foreign Language (EFL) program at King Faisal Schools for grades 1–12. The evaluation is based on an extensive review of instructional materials, including teachers' books, students' books, workbooks, study packets, homework assignments, monthly tests, supplementary materials, and end-of-year student grades. The study describes the textbooks used across all grade levels, outlines program components, and documents the scope and sequence of vocabulary, functions, grammar structures, and language skills (listening, speaking, reading, and writing). It also analyzes the content load, instructional progression, and assessment practices by examining test formats and item types for each grade. The study adopts a descriptive analytical approach, systematically detailing curriculum content, skill coverage, and assessment design to evaluate the strengths and weaknesses of the EFL program across the twelve grade levels.

Cluster 6: Digital Assessment Tools

15) Creating and Sharing iRubrics Using RCampus (Al-Jarf, 2010)

The study demonstrates how EFL college instructors and students can design, use, and share digital rubrics through the iRubric tool in the RCampus learning management system. It explains what rubrics are, distinguishes between holistic and analytic types, and emphasizes the importance of selecting the appropriate scoring method before designing a rubric. The presentation introduces iRubric as a comprehensive platform that allows teachers to define skills, mastery levels, and scoring criteria, attach rubrics to coursework, and use them for grading and student self-assessment. It also outlines how rubric scores are automatically calculated and posted to the gradebook. The advantages highlighted include clearer expectations for students, time-saving grading processes, alignment with learning outcomes, and the ability to share, repurpose, and collaboratively assess rubrics through the RCampus gallery. Overall, the article positions iRubric as a powerful digital tool that enhances transparency, consistency, and collaboration in course assessment.

Cluster 7: Program Admission Policies and Staffing Benchmarks in Language and Translation Education

16) Benchmarks for staffing translation departments in Saudi Arabia (Al-Jarf, 2008)

The study examined the urgent need for new admission benchmarks in Saudi colleges of languages and translation following the nationwide shift to an open-admission policy based solely on high-school GPA. The study highlights a paradox in Fall 2007 admissions: although the lowest admitted GPA was exceptionally high (98.3%), only 21.8% of students passed the reading course, and attrition rates reached 20% in Fall 2003 and 30% in Spring 2004, with continuous weekly withdrawals. The paper analyzes enrolment patterns, failure rates, and re-registration cycles to show that high GPA alone is not a reliable predictor of success in language and translation programs. The author argues that the mismatch between admission criteria and actual student performance indicates a structural problem in program entry requirements. The study concludes by recommending the adoption of new, more valid benchmarks for admission to Saudi language schools to improve student preparedness, reduce attrition, and ensure better alignment between student abilities and program demands.

4. Discussion

4.1 *Meta-Conclusion*

Across the seven clusters synthesized in this review, a unifying pattern emerges: evaluation in language and translation education is not a single practice but an interconnected ecosystem shaped by standards, curricula, instructional quality, assessment design, reviewer expertise, and institutional policy. Although each cluster addresses a distinct domain: thesis evaluation, peer reviewing, teacher performance evaluation, curriculum and textbook evaluation, learning outcomes, digital assessment tools, and admission benchmarks, the collective evidence reveals systemic tensions that recur across levels and contexts. These include misalignment between intended standards and actual practice, variability in evaluator preparation, inconsistencies in assessment design, and gaps between institutional policies and student performance realities. Taken together, the corpus demonstrates that meaningful improvement in educational evaluation requires a holistic, multi-layered approach that integrates curriculum reform, assessment literacy, transparent standards, and institutional accountability. The meta-level insight is clear: evaluation is most effective when it is coherent across the system, data-informed, ethically grounded, and supported by continuous professional development and collaborative research. This review therefore, positions evaluation not as a technical procedure but as a strategic, system-wide commitment essential for improving educational quality in local, regional, and global contexts.

4.2 *Meta-Interpretation*

The meta-interpretation phase sought to move beyond the descriptive synthesis of individual studies to generate higher-order insights about the author's 35-year research program. Because the corpus is internally coherent and methodologically aligned, the interpretive process focused on identifying cross-cutting conceptual patterns, recurring evaluative principles, and the evolution of educational concerns across time, contexts, and domains.

Across clusters, the studies collectively reveal a sustained scholarly commitment to diagnosing systemic weaknesses in educational evaluation, whether in graduate theses, graduate programs, peer-reviewing practices, language-learning curricula, textbook design, learning outcomes or teacher performance. Despite differences in setting and population, the studies converge on several meta-themes: the prevalence of inconsistent standards, the impact of institutional practices on educational quality, the consequences of inadequate supervision or instructional design, and the role of assessment practices in shaping student outcomes. These themes recur across higher education, school-level EFL programs, and professional evaluation contexts, suggesting that the challenges identified are structural rather than isolated.

Additionally, the corpus demonstrates a consistent methodological stance: the use of descriptive, criteria-based evaluation to expose gaps between intended standards and actual practice. Whether examining thesis supervision, peer-reviewing, or curriculum coverage, the studies highlight misalignment between policy and implementation, variability in evaluator expertise, and the consequences of insufficient quality assurance mechanisms. This interpretive pattern underscores the author's broader argument that educational systems require clearer standards, more transparent evaluation processes, and stronger alignment between instructional goals and assessment practices.

Taken together, the meta-interpretation reveals a longitudinal research trajectory centered on improving educational accountability and instructional quality. The studies not only document problems but also implicitly map the structural conditions that produce them, such as assessment policies, curriculum design, institutional culture, and evaluator preparedness. By synthesizing these insights across clusters, the corpus offers a comprehensive, multi-level understanding of how evaluation practices shape educational outcomes in Saudi Arabia and international contexts.

4.3 *Cross-Cutting Insights*

Across the 7 thematic clusters, several cross-cutting insights emerge that illuminate the structural, pedagogical, and evaluative dynamics shaping educational practices in Saudi Arabia and international contexts. Although the studies span different domains, graduate theses, graduate programs, peer-reviewing, language and translation courses, teacher evaluation, and curriculum/textbook evaluation, they converge on a shared set of systemic patterns. The first cross-cutting insight concerns the persistent misalignment between intended standards and actual practice**.** Whether in thesis supervision, peer-reviewing procedures, EFL curriculum implementation, or textbook design, the studies consistently reveal gaps between policy expectations and on-the-ground realities. These gaps manifest as inconsistent evaluation criteria, uneven instructional coverage, variable reviewer expertise, and curricular materials that do not fully meet pedagogical or linguistic benchmarks.

A second insight relates to the central role of evaluator preparedness and institutional culture**.** Across contexts, the studies highlight how the quality of educational outcomes is shaped not only by formal criteria but also by the expertise, workload, and professional norms of those responsible for evaluation—supervisors, peer reviewers, instructors, and program administrators. Weak institutional oversight, insufficient training, and unclear expectations repeatedly emerge as factors that undermine evaluation quality.

A third cross-cutting pattern is the impact of assessment practices on learning outcomes**.** Studies on grade inflation, course-grade distributions, and EFL curriculum coverage show that assessment systems often fail to reflect actual student performance or curricular goals. This misalignment contributes to inflated grades, superficial learning, and reduced accountability, echoing similar concerns identified in thesis evaluation and peer-reviewing studies.

A fourth insight is the recurrence of structural constraints—such as time pressure, heavy workloads, limited resources, and inadequate quality-assurance mechanisms—that shape educational processes across all levels. These constraints appear in peer-reviewing delays, insufficient supervision of graduate students, incomplete coverage of reading textbooks, and inconsistent implementation of EFL curricula.

Finally, the corpus reveals a broader, longitudinal insight: educational evaluation practices across domains are interconnected**,** and weaknesses in one area often mirror or reinforce weaknesses in another. For example, inadequate training in research methods at the undergraduate or MA level later manifests as weak thesis writing, which in turn burdens peer reviewers and undermines publication quality. Similarly, gaps in school-level EFL curricula echo later in university-level language and translation performance.

Taken together, these cross-cutting insights demonstrate that the challenges documented across the corpus are not isolated phenomena but part of a larger ecosystem of evaluation practices. The studies collectively point to the need for clearer standards, stronger evaluator preparation, more coherent assessment systems, and sustained institutional commitment to quality assurance across all levels of education.

4.4 *Implications*

The cross-cluster synthesis yields several important implications for educational policy, instructional practice, evaluator preparation, and future research. Because the corpus spans multiple levels of the educational system as school curricula, graduate programs, peer-reviewing, and teacher evaluation, the implications extend across institutional boundaries and highlight systemic needs. The first implication concerns the urgent need for clearer, standardized evaluation criteria across all educational domains. The studies consistently reveal that inconsistent or poorly defined standards undermine the quality of thesis supervision, peer-reviewing, curriculum implementation, and assessment practices. Establishing transparent, discipline-appropriate benchmarks would enhance fairness, reduce variability in evaluator judgments, and improve the reliability of educational outcomes.

A second implication relates to evaluator training and professional development. Across contexts, the findings show that supervisors, peer reviewers, instructors, and program evaluators often lack adequate preparation for their evaluative roles. Structured training in assessment literacy, research-methodology evaluation, and curriculum alignment would strengthen the quality of decision-making and reduce the systemic weaknesses documented across the corpus.

A third implication concerns assessment reform. Studies on grade inflation, course-grade distributions, and textbook coverage demonstrate that assessment practices frequently fail to reflect actual learning or curricular goals. Institutions may need to revise grading policies, diversify assessment formats, and implement monitoring systems to ensure that grades accurately represent student performance and that instructional coverage aligns with course objectives.

A fourth implication is the importance of curriculum coherence and instructional alignment. The analyses of EFL programs, reading-textbook coverage, and CEFR-aligned materials highlight the need for curricula that are developmentally sequenced, pedagogically balanced, and realistically paced. Ensuring alignment between textbooks, instructional practices, and assessment systems would improve learning outcomes and reduce the instructional gaps identified in the studies.

A fifth implication concerns institutional accountability and quality assurance. Weak oversight mechanisms—whether in peer-reviewing, thesis evaluation, or program implementation—allow inconsistencies to persist. Strengthening institutional policies, monitoring procedures, and feedback systems would enhance transparency and support continuous improvement across educational levels.

Finally, the corpus underscores the need for longitudinal, system-wide evaluation research. The recurring patterns identified across 35 years suggest that many challenges are structural rather than isolated. Future research should therefore adopt multi-level, longitudinal designs that examine how early educational practices (e.g., school-level EFL instruction) shape later outcomes in university performance, research quality, and professional evaluation.

Taken together, these implications point to the necessity of coordinated reform efforts that integrate curriculum design, evaluator preparation, assessment policy, and institutional quality assurance. The corpus demonstrates that meaningful improvement in educational outcomes requires systemic, rather than piecemeal, change.

4.5 *Positioning This ISR Within the Global Educational Evaluation SR/MA Research*

This systematic review occupies a distinctive position within the global body of SR/MA research on educational evaluation. While international reviews have traditionally focused on large-scale assessment systems, national curriculum reforms, teacher effectiveness, literacy development, and program evaluation in primary and secondary education, very few have examined an author-bounded, longitudinal research program spanning multiple educational levels and evaluation domains. In this respect, the present ISR contributes a unique methodological and conceptual perspective to the global literature.

Globally, SRs in education tend to synthesize heterogeneous studies conducted by multiple researchers across diverse contexts. By contrast, this ISR synthesizes a coherent, 35-year corpus produced by a single scholar, enabling a level of internal consistency and longitudinal insight that is rarely achievable in conventional SR/MA designs. This approach aligns with emerging trends in research-program analysis, where scholars examine the evolution of a single researcher's contributions to trace conceptual development, methodological continuity, and thematic progression over time.

Furthermore, this ISR expands the scope of global educational evaluation research by integrating domains that are often treated separately in international SRs, such as thesis evaluation, peer-reviewing practices, curriculum and textbook evaluation, grade-inflation studies, and teacher-performance assessment, admission benchmarks, and digital assessment tools. In doing so, it highlights the interconnectedness of evaluation practices across the educational ecosystem, offering a multi-level perspective that complements global SRs focused on isolated components of educational evaluation.

This ISR also contributes to the international literature by foregrounding Arab and Saudi educational contexts, which remain underrepresented in global SR/MA research. By synthesizing studies conducted across local, regional, and international settings, this ISR provides a rare longitudinal account of educational evaluation practices in the Arab world, thereby filling a geographic and cultural gap in the global literature.

Finally, the methodological orientation of this ISR, emphasizing descriptive, criteria-based evaluation, adds a valuable dimension to global SR/MA traditions that often prioritize experimental or quasi-experimental designs. The corpus demonstrates how descriptive evaluation research can generate actionable insights into systemic weaknesses, institutional practices, and curriculum alignment, offering a model for integrating qualitative and descriptive evidence into broader educational evaluation frameworks.

Taken together, this ISR positions itself as a novel contribution to global educational evaluation research: one that bridges methodological traditions, expands geographic representation, and demonstrates the value of synthesizing a coherent, author-bounded research program to illuminate long-term patterns in educational evaluation.

4.6 *Limitations of This ISR*

Despite its methodological coherence and longitudinal depth, this ISR has several limitations that should be acknowledged when interpreting its findings. First, the corpus is author-bounded, consisting exclusively of studies conducted by a single researcher over a 35-year period. While this provides exceptional internal consistency, it also limits the diversity of methodological perspectives, theoretical frameworks, and evaluative approaches typically found in multi-author SR/MA research. As a result, the synthesis reflects the evolution of one research program rather than the full spectrum of global scholarship on educational evaluation.

Second, the included studies are predominantly descriptive and qualitative, relying on frequency counts, document analysis, and criteria-based evaluation rather than experimental or quasi-experimental designs. This limits the ability to calculate effect sizes, conduct statistical meta-analysis, or generalize findings beyond the contexts examined. This ISR therefore offers conceptual and thematic insights rather than causal claims or quantitative estimates of impact.

Third, the corpus spans a variety of evaluation domains, thesis evaluation, peer-reviewing, curriculum and textbook evaluation, learning outcomes and course grades, teacher performance, admission benchmarks, and assessment tools which, vary in scope, population, and methodological emphasis even though they are conceptually connected. Although thematic clustering mitigates this heterogeneity, the breadth of topics may still constrain the depth of synthesis within any single domain.

Fourth, several studies rely on convenience samples, such as voluntary survey responses or institution-specific datasets. These sampling approaches may introduce self-selection bias and limit the representativeness of findings. Similarly, some studies draw on institutional documents or course-grade distributions that may not reflect broader national or international patterns.

Fifth, because the ISR synthesizes studies conducted across different decades, changes in educational policy, institutional structures, and technological environments may influence the comparability of findings over time. Earlier studies may reflect conditions that no longer exist, while more recent studies may capture emerging practices not present in earlier phases of the research program.

Finally, the ISR is limited by the absence of external validation. Since the corpus is author-bounded, triangulation with independent studies, alternative evaluation frameworks, or cross-institutional datasets was not possible within the scope of this ISR. Future SRs may incorporate multi-author, multi-context evidence to help situate findings within a broader empirical framework.

Taken together, these limitations do not diminish the value of the synthesis but rather contextualize its contributions. This ISR offers a unique longitudinal perspective on educational evaluation, while acknowledging the boundaries within which its insights should be interpreted.

4.7 *Future Research Directions*

The ISR of 35-year author-bounded corpus highlights several directions for future research in educational evaluation, curriculum studies, and assessment practices as follows:

The first direction involves conducting multi-author, multi-institutional studies that examine the same evaluation domains explored in this ISR, thesis supervision, peer-reviewing, graduate program evaluation, curriculum evaluation, textbook evaluation, grade outcomes, teacher performance and assessment tools. Comparative studies across institutions, regions, and countries would help determine whether the patterns identified here are locally specific or globally recurrent.

A second direction concerns the need for mixed-methods and longitudinal designs. Future research could incorporate experimental, quasi-experimental, or longitudinal tracking approaches to examine causal relationships between evaluation practices and educational outcomes. Such designs would strengthen the empirical foundations of evaluation research and allow for more robust generalization.

A third direction is the development of validated evaluation frameworks and assessment tools. The recurring inconsistent standards, variable evaluator expertise, and misalignment between curriculum and assessment across the corpus underscore the need for standardized assessment tools that can be used across institutions. Future research could focus on designing, piloting, and validating such tools for thesis evaluation, peer-reviewing, curriculum and teacher performance assessment.

A fourth direction involves expanding research into underexplored educational contexts, particularly within the Arab world. Many domains, such as digital peer-reviewing, online thesis supervision, AI-supported assessment, and CEFR-aligned curriculum implementation, are still insufficiently studied in Arab educational systems. Future studies could address these gaps and contribute to a more globally representative evidence base.

A fifth direction concerns student-level learning analytics. Studies on grade inflation, course-grade distributions, and textbook coverage suggest the need for more detailed analyses of student learning outcomes. Future research could employ learning analytics, classroom observations, and performance-tracking systems to examine how curriculum design, instructional practices, and assessment policies shape student outcomes over time.

Finally, future research should explore institutional reform and policy implementation. The corpus highlights systemic issues that require coordinated institutional responses. Research examining how universities, schools, and ministries implement evaluation reforms and how these reforms influence educational quality would provide valuable insights into the mechanisms of sustainable change.

Taken together, these directions point toward a research agenda that is broader, more collaborative, and more methodologically diverse. Building on the foundations of this corpus, future studies can contribute to a more comprehensive and globally informed understanding of educational evaluation.

5. Recommendations

Based on the cross-cluster synthesis and the meta-interpretation of 35 years of research, this ISR offers the following recommendations to educators, evaluators, policymakers, institutions, and researchers and for improving interconnected domains of thesis evaluation, peer-reviewing, curriculum design, textbook development, assessment practices, teacher performance, admission benchmarks and assessment tools:

- Strengthening evaluation systems begins with ensuring that standards are transparent, accessible, and consistently applied across institutions. This requires aligning evaluation criteria with internationally recognized benchmarks such as CEFR for language curricula, APA/MLA for academic writing, and COPE for peer-reviewing ethics. Transparency is further reinforced by documenting evaluation procedures, reviewer guidelines, and program outcomes, allowing institutions to maintain consistency, reduce ambiguity, and promote fairness across all levels of academic evaluation.

- High-quality educational programs depend on strong alignment between curricula, textbooks, instructional practices, and assessment methods. Institutions are encouraged to adopt evidence-based textbook evaluation frameworks to guide material selection and development, ensuring that content is appropriate in scope, sequence, and skill progression. Regular curriculum audits help identify gaps in pacing, content load, and learning trajectories. Additionally, collecting and analyzing data on student performance, curriculum coverage, reviewer feedback, and program outcomes enables continuous improvement and supports data-driven revisions to curricula, teacher training, and institutional policies.

- Improving assessment quality requires diversifying evaluation formats to include performance-based tasks, portfolios, and criterion-referenced assessments that more accurately capture student learning. Institutions should regularly review grading policies to address grade inflation and ensure that grades reflect student performance. Establishing monitoring systems to track grade distributions, assessment alignment, and instructional coverage helps maintain academic integrity and supports evidence-based decision-making at the program and institutional levels.

- Robust quality assurance depends on establishing internal review committees that oversee thesis evaluation, peer-reviewing processes, and program implementation. Effective systems also incorporate feedback loops that allow students, instructors, and reviewers to report inconsistencies or systemic issues. Strengthening evaluator expertise is essential; therefore, institutions should provide structured training in assessment literacy, evaluation methodology, curriculum alignment, research-methodology evaluation, and ethical reviewing practices. As academic ecosystems evolve, emerging areas such as digital peer reviewing, AI-supported assessment, online thesis supervision, and remote curriculum delivery warrant further investigation and integration into institutional evaluation frameworks.

- Advancing evaluation practices requires sustained collaboration among researchers, ministries, and educational institutions to develop shared frameworks and unified standards. Multi-institutional studies are essential for validating findings across diverse contexts and reducing reliance on single-site evidence. Comparative research that examines evaluation practices across Arab and international settings can illuminate regional strengths and gaps, while targeted studies in underrepresented Arab regions and institutions help build a more inclusive and comprehensive understanding of evaluation systems across the Arab world.

6. Conclusion

This ISR synthesized a 35-year, author-bounded research program that spans multiple domains of educational evaluation, including thesis supervision, graduate-program quality, peer-reviewing practices, curriculum and textbook evaluation, learning outcomes, teacher performance and evaluation tools. By integrating 16 methodologically coherent studies, the ISR provides a rare longitudinal perspective on how evaluation practices evolve across time, institutions, and educational levels and domains. The ISR revealed several recurring patterns: persistent misalignment between intended standards and actual practice; variability in evaluator expertise; structural constraints that shape instructional and evaluative processes; and the far-reaching impact of assessment policies on student learning and institutional accountability. These patterns emerged consistently across school-level EFL programs, university-level language and translation courses, graduate-level research supervision, and professional peer-reviewing contexts, demonstrating that the challenges identified are systemic rather than isolated.

The ISR also highlights the value of descriptive, criteria-based evaluation as a methodological approach capable of uncovering structural weaknesses that may be overlooked in experimental or quantitative research. By examining curricula, textbooks, assessment systems, and evaluator practices, the corpus offers a holistic understanding of how educational quality is produced and compromised within real institutional settings.

Positioned within the global educational-evaluation literature, this ISR contributes a distinctive perspective by foregrounding Arab and Saudi contexts, integrating multiple evaluation domains, and synthesizing a coherent research trajectory rather than a heterogeneous body of unrelated studies. It fills a notable gap in international SR/MA scholarship, which rarely includes longitudinal, author-bounded analyses or comprehensive evaluations of educational ecosystems in the Arab world.

The implications of this review point toward the need for clearer evaluation standards, stronger evaluator preparation, improved curriculum alignment, and more robust institutional quality-assurance mechanisms. Future research should expand beyond single-author corpora to include mixed-methods, cross-national studies and multi-institutional that validate and extend the insights generated here.

Overall, this ISR demonstrates that meaningful improvement in educational evaluation requires coordinated, system-wide reform. By tracing the evolution of evaluation practices across decades and domains, the ISR provides a foundation for future scholarship and policy development aimed at enhancing educational quality, accountability, and equity across local, regional, and global contexts.

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

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

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). 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

[4] Al-Jarf, R. (2026d). 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

[5] Al-Jarf, R. (2026e). 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

[6] Al-Jarf, R. (2026f). 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

[7] Al-Jarf, R. (2026g). 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

[8] Al-Jarf, R. (2026h). 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

[9] Al-Jarf, R. (2026i). 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

[10] Al-Jarf, R. (2026j). 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

[11] Al-Jarf, R. (2026k). 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

[12] Al-Jarf, R. (2026l). 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

[13] Al-Jarf, R. (2026m). 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

[14] Al-Jarf, R. (2026n). 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

[15] Al-Jarf, R. (2026o). 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

[16] Al-Jarf, R. (2026p). 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

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

[18] Al-Jarf, R. (2023b). Preparing high schools students for the university and life after graduation. *Eurasian Arabic Studies, 6*(3), 93-116. DOI: 10.26907/2619-1261.2023.6.3.93-116. Google Scholar

[19] Al-Jarf, R. (2023c). 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

[20] Al-Jarf, R. (2022a). 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

[21] Al-Jarf, R. (2022b). 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

[22] Al-Jarf, R. (2022c). 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

[23] Al-Jarf, R. (2022d). 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

[24] Al-Jarf, R. (2022e). 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

[25] 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

[26] 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

[27] Al-Jarf, R. (2021c). 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

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

[29] Al-Jarf, R. (2021e). *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

[30] 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

[31] 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

[32] 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

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

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

[35] 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

[36] 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

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

[38] 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

[39] Al-Jarf, R. (2013). *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

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

[41] 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

[42] 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

[43] 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

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

[45] Al-Jarf, R. (2009). *How to prepare English language tests*. COLT Symposium series. Riyadh, Saudi Arabia. December 26.

[46] Al-Jarf, R. (2008). *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, 2008. ERIC ED611785, Google Scholar

[47] Al-Jarf, R. (2008a). *MA and Ph.D. Thesis evaluation problems and proposed solutions*. Symposium on Peer-reviewing. Imam Mohammad Bin Saud University. Riyadh, Saudi Arabia. November 19-20. Google Scholar https://www.researchgate.net/publication/355484478

[48] Al-Jarf, R. (2008b). *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

[49] 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

[50] Al-Jarf, R. (2007b). *Assessing students' reading competencies: Setting global standards*. AEJ Global Congress. Seoul, Korea. May 25-26.

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

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

[53] 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

[54] Al-Jarf, R. (2002a). *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

[55] Al-Jarf, R. (2002b). *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

[56] 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

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

[58] 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

[59] 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

[60] 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

[61] Abbas, M. & Hameed, S. (2022). A systematic review of deep learning based online exam proctoring systems for abnormal student behaviour detection. International *Journal of Scientific Research in Science, Engineering and Technology*, 9(4), 192-209.

[62] Agir, N., et al. (2023). Outcome-based assessment in the evaluation of education programs through a systematic literature review. *International Journal of Academic Research in Progressive Education and Development, 12*(2), 2483-2497.

[63] Ahmed, Y., Kanyengo, C. & Akakandelwa, A. (2010). Mapping Postgraduate Research at the University of Zambia: a review of dissertations for the Master of Medicine Programme. Medical journal of Zambia, 37(2), 52.

[64] Al-Alawi, R., & Alexander, G. (2020). Systematic review of program evaluation in baccalaureate nursing programs. Journal of Professional Nursing, 36(4), 236-244.

[65] Albayrak, E. (2022). A review of the studies conducted on online exams in Turkey from the millennium to the coronavirus period. *Instructional Technology and Lifelong Learning, 3*(2), 207-224.

[66] Alzahrani, M. & Nor, F. (2021). A systematic review to Identify EFL teachers' training needs for professional development programs. *International journal of academic research in progressive education and development, 10*(3).

[67] Babik, D., et al. (2024). A systematic review of educational online peer-review and assessment systems: Charting the landscape. *Educational technology research and development, 72*(3), 1653-1689.

[68] Baghian, N., et al. (2019). Evaluation of students' mental and social health promotion educational programs: A systematic review. *Journal of education and health promotion, 8*(1), 258.

[69] Bengtsson, L. (2019). Take-home exams in higher education: A systematic review. *Education Sciences, 9*(4), 267.

[70] Bervell, B., et al. (2025). Web-based examinations in higher education (WEBiHE) institutions in the Sub-Saharan Africa region: a systematic review of 2013–2024 literature. *Cogent Education, 12*(1), 2519565.

[71] Braun, M. (2024). A literature review of online exams in HE in Physics and Maths. *New Directions in the Teaching of Natural Sciences,* 19.

[72] Casey, S. (2015). *Evaluations of reproductive health programs in humanitarian settings: a systematic review*. Conflict and health, 9 (Suppl 1), S1.

[73] Chong, S. & Lin, T. (2024). Feedback practices in journal peer-review: a systematic literature review. *Assessment & Evaluation in Higher Education, 49*(1), 1-12.

[74] Dayananda, D., et al. (2021). A systematic literature review on online exams in COVID-19 pandemic: Assessment methods, students' preferences, dishonest behaviors and challenges in online exams. 2021 *From Innovation to Impact (FITI), 1*, 1-6.

[75] Farahsa, S., & Tabrizi, J. (2015). How evaluation and audit is implemented in educational organizations? A systematic review. *Research and Development in Medical Education, 4*(1), 3-16.

[76] Fatima, T., Azam, F., & Muzaffar, A. (2022). *A systematic review on fully automated online exam proctoring approaches.* In 2022 24th International multitopic conference (INMIC) (pp. 1-5). IEEE.

[77] Felthun, J. et al. (2021). Assessment methods and the validity and reliability of measurement tools in online objective structured clinical examinations: a systematic scoping review. *Journal of Educational Evaluation for Health Professions, 18*.

[78] Glonti, K., et al. (2019). A scoping review on the roles and tasks of peer reviewers in the manuscript review process in biomedical journals. *BMC medicine, 17*(1), 118.

[79] Horne, E. & Sandmann, L. (2012). Current trends in systematic program evaluation of online graduate nursing education: An integrative literature review. *Journal of Nursing Education, 51*(10), 570-578.

[80] Hos, R., & Topal, H. (2013). The current status of English as a Foreign Language (EFL) teachers' professional development in Turkey: A systematic review of literature. *The Anthropologist, 16*(1-2), 293-305.

[81] İpek, Ö. (2022). A Systematic review of Program Evaluation Studies in EFL: The Turkish Case. *Dil ve Edebiyat Araştırmaları,* 25, 199-217.

[82] Iqbal, Z., et al. (2021). A comparative analysis of the efficacy of three program-evaluation models–A review on their implication in educational programs. *Humanities & Social Sciences Reviews, 9*(3), 326-336.

[83] Kaddoura, S., Popescu, D. & Hemanth, J. (2022). A systematic review on machine learning models for online learning and examination systems. *PeerJ Computer Science, 8*, e986.

[84] Kuleva, M., & Miladinov, O. (2024). *Exploring the efficacy of online proctoring in online examinations: A comprehensive review.* In Environment. Technology. Resources. Proceedings of the International Scientific and Practical Conference (Vol. 2, pp. 192-196).

[85] Lasker, S. (2018). Peer review system: a systematic review. *Bangladesh Journal of Bioethics, 9*(1), 13-23.

[86] Lefebvre, C., & Duffy, S. (2021). Peer review of searches for studies for health technology assessments, systematic reviews, and other evidence syntheses. *International Journal of Technology Assessment in Health Care, 37*(1), e64.

[87] Malhotra, M., & Chhabra, I. (2026). Ensuring academic integrity through automated online exam proctoring a decade long systematic review. *Discover Education*.

[88] Muzaffar, A. W., et al. (2021). A systematic review of online exams solutions in e-learning: Techniques, tools, and global adoption. *IEEE Access, 9*, 32689-32712.

[89] Năznean, A. (2021). Cheating during online examinations-literature review. *Journal of Pedagogy*, 2.

[90] Nicoll, P., et al. (2018). Evaluation of technology-enhanced learning programs for health care professionals: systematic review. *Journal of medical Internet research, 20*(4), e9085.

[91] Nouraey, P., et (2020). Educational program and curriculum evaluation models: a mini systematic review of the recent trends. *Universal Journal of Educational Research, 8*(9), 4048-4055.

[92] Oliveira, C., De Man, A., & Guerreiro, S. (2015). Tourism research: A systematic review of knowledge and cross cultural evaluation of doctoral theses. *Tourism & Management Studies, 11*(1), 111-119.

[93] Öncül, B. (2021). Dealing with cheating in online exams: A systematic review of proctored and non-proctored exams. *International technology and education journal, 5*(2), 45-54.

[94] Osman, K. (2023). A systematic literature review of authentic assessment in K-12 ESL/EFL education. *Malaysian Journal of Social Sciences and Humanities (MJSSH), 8*(5), e002303-e002303.

[95] Parylo, O. (2012). Evaluation of educational administration: A decade review of research (2001–2010). *Studies in Educational Evaluation, 38*(3-4), 73-83.

[96] Puspawati, I., Khansa, M., & Widiati, U. (2024). Developing EFL teachers' language assessment literacy: A systematic literature review on teacher training programs. *TESL-EJ, 28*(2), n2.

[97] Ramírez de Dampierre, M., et al. (2024). Evaluation of the implementation of Project-Based-Learning in engineering programs: A review of the literature. *Education Sciences, 14*(10), 1107.

[98] Saleh, G., Tharwat, G., & Gamalel-Din, S. (2023). A systematic survey on examinees identity authentication in online distant exams. *Journal of Al-Azhar University Engineering Sector, 18*(66), 129-151.

[99] Shunko, A. (2025). Open book exams in higher education: A systematic review. *Pedagogy and Psychology, 63*(2), 5-20.

[100] Sizo, A., et al. (2025). Defining quality in peer review reports: a scoping review. *Knowledge and Information Systems, 67*(8), 6413-6460.

[101] Spatıotı, A., Kazanıdıs, I., & Pange, J. (2023). Educational design and evaluation models of the learning effectiveness in e-learning process: a systematic review. *Turkish Online Journal of Distance Education, 24*(4), 318-347.
[102] Tas, I. D., & Duman, S. N. (2021). A systematic review of Postgraduate Theses on Curriculum Evaluation. International *Journal of Curriculum and Instructional Studies, 11*(1), 43-64.
[103] Tian, J., et al. (2007). A systematic review of evaluation in formal continuing medical education. *Journal of continuing education in the health professions, 27*(1), 16-27.
[104] Todres, R., & Bunston, T. (2009). Parent education program evaluation: A review of the literature. *Canadian Journal of Community Mental Health, 12*(1), 225-257.
[105] Topuz, A. C., & Kinshuk (2024). Students' acceptance of and preferences regarding online exams: a systematic literature review. *Educational technology research and development, 72*(2), 1111-1151.
[106] Ullah, H., et al. (2024). Curriculum and program evaluation in medical education: a short systematic literature review. *Annals of Medicine and Surgery, 86*(10), 5988-5994.
[107] Vidair, H., et al. (2019). A systematic review of research on dissertations in health service psychology programs. *Training and Education in Professional Psychology, 13*(4), 287.
[108] Walser, T. & Trevisan, M. (2016). Evaluability assessment thesis and dissertation studies in graduate professional degree programs: review and recommendations. *American Journal of Evaluation, 37*(1), 118-138.
[109] Wannas, A. & AbdelMohsen, M. (2025). A systematic review of EFL online assessment in higher education: effectiveness, attitudes, and challenges. *Knowledge Management & E-Learning, 17*(3), 435-453.
[110] Watts, L., et al. (2017). Qualitative evaluation methods in ethics education: A systematic review and analysis of best practices. *Accountability in Research, 24*(4), 225-242.

Over the years, high teacher turnover has become a serious issue around the world. The problem has been linked to poor workforce