Language Assessment in Pakistan: A PRISMA-Guided Systematic Review of Teachers' Perceptions, Practices and Challenges

[1]Syed Najmul Hassan and [2]Rashida Imran

[1]PhD Scholar, Department of English and Applied Linguistics

Allama Iqbal Open University, Islamabad, Pakistan.

[2]Assistant Professor, Department of English and Applied Linguistics

Allama Iqbal Open University, Islamabad, Pakistan. Email: rashida.imran@aiou.edu.pk

Corresponding author:

Syed Najmul Hassan

najamshahsyed@gmail.com

Abstract

Language assessment literacy (LAL) has emerged as a crucial competence for English language teachers worldwide. In Pakistan, an exam-driven system and resource constraints create distinctive conditions that merit systematic synthesis. This study systematically reviews how English language teachers in Pakistan perceive and practice assessment and the challenges they face. To our knowledge, this is the first PRISMA-guided systematic synthesis of research on English language teachers' assessment perceptions, practices, and challenges in Pakistan. Following PRISMA 2020 guidelines, we searched Scopus, ERIC, Web of Science, Google Scholar, and Pakistani national journals for English language studies published between 1989 and 2024. After duplicate removal, 210 records were screened; 58 full texts were assessed; and 32 studies met the inclusion criteria. Studies were coded against the Teacher Assessment Literacy in Practice (TALiP) model (Xu & Brown, 2016) and teacher cognition framework (Borg, 2015). Findings indicate that teachers predominantly equated assessment with summative examinations, while formative and authentic assessment remained limited. Practices often reproduced high-stakes examination formats, with insufficient attention to speaking and listening skills. Limited professional preparation for LAL, strong washback from board examinations, and persistent concerns about fairness and validity constrained pedagogical innovation. Although portfolios and peer assessment appeared in some better-resourced settings, wider adoption remained limited. Overall, Pakistani teachers' assessment practices remain exam-centric and shaped by systemic constraints and entrenched beliefs. Strengthening LAL therefore requires context-sensitive professional development alongside examination reform that promotes communicative competence, fairness and meaningful learning.

Keywords: *Pakistan; language assessment; assessment literacy; teacher cognition; TALiP; washback; PRISMA; English language teaching*

1. Introduction

Language assessment plays a pivotal role in shaping both teaching and learning. It influences curricular implementation, instructional practices, and students' academic and professional trajectories. Globally, the discourse on Language Assessment Literacy (LAL) has emphasized teachers' knowledge, skills, and beliefs in designing, implementing, and interpreting assessments (Inbar-Lourie, 2008; Fulcher, 2012; Xu & Brown, 2016). Language assessment literacy is a wider construct which denotes language teachers' knowledge, skills and competence to design, administer, interpret and utilize language assessments effectively (Inbar-Lourie, 2013). Unlike language assessment, which focuses on measuring learners' abilities, LAL emphasizes the knowledge of assessment principles among educators, test designers, and policymakers. According to Taylor (2013) LAL consists of awareness of validity, reliability, fairness, and ethical considerations in evaluation.

Harding and Kremmel (2016) say that LAL is essential for teachers because it enables them to make informed decisions about the types of assessments they use, interpret assessment results accurately, and provide meaningful feedback to learners. Teachers with strong LAL are more likely to integrate formative assessment strategies, use assessments to support learning, and align their evaluation methods with language learning objectives (Scarino, 2013).

Different studies (Fulcher, 2012; Inbar-Lourie, 2013a; Giraldo, 2018b) suggest that teachers are required to possess a substantial quantity of expertise in language evaluation, including information, abilities, and principles. While several studies indicate that teachers in preparation and in the field need training in all the areas of language assessment (Fulcher, 2012; Vogt & Tsagari, 2014; Giraldo & Murcia, 2018), recent investigations show the influence of institutional contexts on the LAL of the teachers (Scarino, 2013; Hill, 2017).

In Pakistan, however, the assessment landscape is marked by exam-centric systems, large class sizes, and limited teacher preparation, making it a particularly complex context for understanding LAL.

English in Pakistan occupies a dual role, as a medium of instruction in many schools and universities and as a gatekeeping language for higher education, employment, and mobility (Mahboob, 2003; Rahman, 2002). Consequently, high-stakes examinations in English exert

profound washback effects, narrowing classroom practices and reinforcing rote-based learning (Rehmani, 2003; Malik & Qureshi, 2018). Teachers often act less as facilitators of learning and more as exam-preparation coaches. Despite awareness of broader assessment principles such as validity, reliability, authenticity, and fairness (Messick, 1989; McNamara, 2000), the institutional ecology frequently restricts teachers to traditional written tests, marginalizing formative and alternative assessments.

1.1 Statement of the Problem

While numerous individual studies in Pakistan have examined teachers' assessment perceptions, classroom practices, and assessment-related challenges (e.g., Hidri et al., 2024; Asif & Naz, 2021; Shahzad et al., 2022), the evidence has not been systematically synthesized across these dimensions. Global reviews of language assessment literacy (Vogt & Tsagari, 2014; Tsagari & Vogt, 2017) provide an important comparative foundation but do not synthesize the Pakistani evidence base. The present review therefore addresses this gap by bringing together studies of teachers' assessment conceptions, practices, use of assessment evidence, professional preparation and systemic constraints. The claim that this is the first synthesis is based on the review's database search and screening of the literature on Pakistani language assessment; however, because no separate registered scoping review of prior reviews was conducted, the novelty claim is appropriately qualified as to our knowledge.

1.2 Conceptual Framework

This review draws on two complementary frameworks:

- ***Teacher Assessment Literacy in Practice (TALiP)*** (Xu & Brown, 2016), which conceptualizes assessment literacy across four domains, i.e., knowledge, purposes, design/implementation, and interpretation/use.
- ***Teacher cognition*** (Borg, 2015), which explains how teachers' beliefs and contextual pressures mediate their enacted assessment practices.

By integrating TALiP with teacher cognition, this review not only describes what Pakistani teachers know and do about assessment but also explains *why* their practices persist in exam-dominated systems.

1.3 Review Objectives and Research Questions

Guided by PRISMA 2020 principles (Page et al., 2021), this systematic review synthesizes evidence on English language teachers' assessment perceptions, practices, use of assessment evidence, professional preparation, and systemic constraints in Pakistan. The objectives and research questions are deliberately aligned: each objective specifies the synthesized evidence, while the corresponding question states the analytical issue addressed in the narrative thematic synthesis.

Objective 1: To identify and synthesize how Pakistani English language teachers conceptualize the purposes of classroom assessment.

RQ1: How do Pakistani English language teachers conceptualize the purposes of assessment?

Objective 2: To map and analyze the classroom assessment practices and tools reported in research on English language teaching in Pakistan.

RQ2: What classroom assessment practices and tools are reported in the literature?

Objective 3: To examine how teachers interpret and use assessment evidence for instructional decision-making.

RQ3: How do teachers interpret and use assessment evidence?

Objective 4: To evaluate the status and scope of teacher education and professional development in Language Assessment Literacy (LAL) in Pakistan.

RQ4: What is the status of teacher education and professional development in LAL in Pakistan?

Objective 5: To analyze systemic challenges, including washback, fairness, and practicality, that shape teachers' assessment practices.

RQ5: What systemic challenges, including washback, fairness, and practicality, shape assessment practices?

2. Methods

2.1 Protocol and Reporting Standards

This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement (Page et al., 2021) to ensure methodological transparency and replicability. The protocol was designed to map evidence on Pakistani English language teachers' perceptions, practices, and challenges in language assessment. While no prior registration was made (e.g., PROSPERO), the review process was informed by PRISMA guidelines and adapted for conceptual and educational research.

2.2 Eligibility Criteria

Studies were included if they met the following criteria:

- ***Population:*** Pre-service or in-service English language teachers or English language teacher educators working in Pakistan.
- ***Concept:*** Language assessment or assessment literacy, including perceptions, practices, training, washback, fairness, and challenges.
- ***Context:*** Pakistani school, college, or higher education settings. Multi-country studies were included only if they reported **disaggregated data** for Pakistan.
- ***Study types:*** Empirical (qualitative, quantitative, or mixed-methods), conceptual/theoretical, or policy analyses published in peer-reviewed journals, edited books, or rigorously curated reports.
- ***Language:*** English.
- ***Timeframe:*** January 1989 – March 2024 (starting from Messick's *validity* framework).

Exclusion criteria were: (a) opinion/editorial pieces without empirical or theoretical grounding, (b) psychometric-only studies without implications for teachers, and (c) non-English publications.

2.3 Information Sources and Search Strategy

Searches were conducted between August and October 2025 across international and national databases:

- Scopus
- Web of Science (WoS)
- ERIC (Education Resources Information Center)
- Google Scholar
- Pakistani-based journals and repositories

A comprehensive search strategy was designed to capture both international and local publications. Boolean operators and truncation were applied. Example search string (ERIC, September 2025):

("language assessment" OR "assessment literacy" OR "formative assessment" OR "washback")

AND (Pakistan OR Pakistani)

AND (teacher OR instructor OR educator)

2.4 Study Selection

All identified records were exported to EndNote, and duplicates were removed. Screening proceeded in three stages:

1. Title/abstract screening against inclusion criteria.
2. Full-text review for eligibility.
3. Final inclusion based on alignment with the TALiP and teacher cognition frameworks.

Two reviewers independently screened all titles/abstracts and full texts. Disagreements were resolved through discussion.

2.5 PRISMA Flow Diagram

The selection process is summarized in Figure 1 using the PRISMA 2020 flow structure. A total of 234 records were identified, 24 duplicates were removed, and 210 records underwent title/abstract screening. Of these, 152 were excluded, leaving 58 full texts for eligibility assessment. Twenty-six full texts were excluded, resulting in 32 studies in qualitative synthesis. The same counts are used consistently throughout the manuscript.

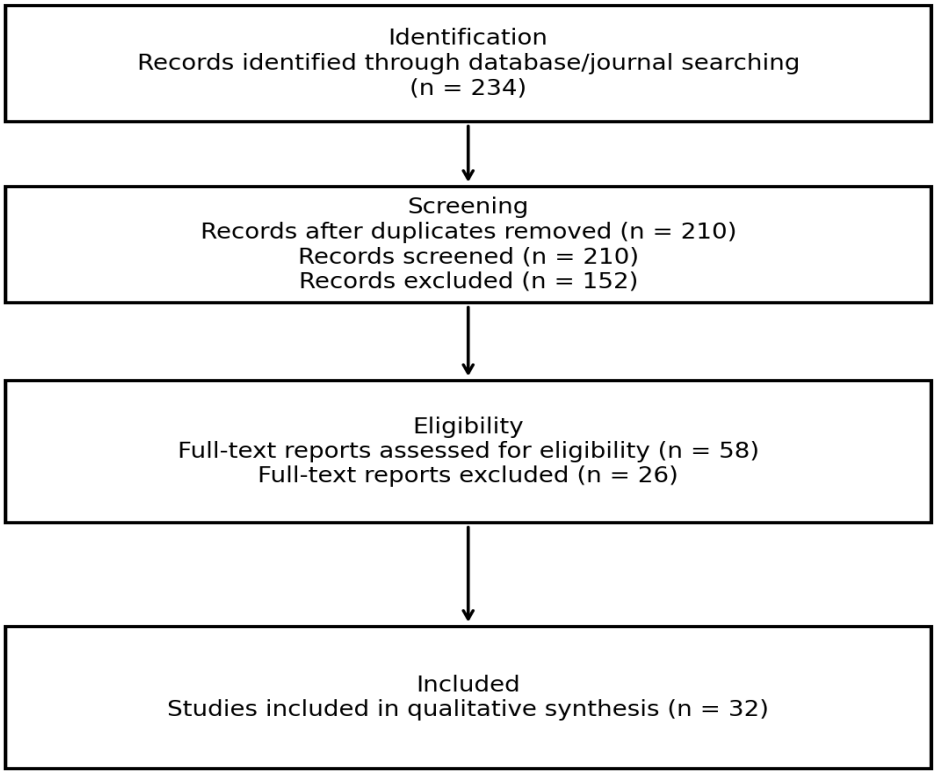


Figure No.1. PRISMA 2020 Flow Diagram

2.6 Data Extraction

A standardized data extraction form was developed. Key information included:

- Bibliographic details (author, year, publication type)
- Study context (level, province, institution type)
- Participants (sample size, teacher type)
- Methodology (qualitative/quantitative/mixed; instruments used)

Two reviewers independently screened titles/abstracts and full texts against the eligibility criteria. Disagreements were resolved through discussion. No formal inter-rater agreement statistic (e.g., Cohen's kappa) was calculated; this is reported explicitly as a methodological deviation and limitation rather than being implied as evidence of agreement. Future updates of the review should record independent decisions and calculate inter-rater agreement before consensus resolution.

- Main findings (perceptions, practices, training, challenges, washback)
- Alignment with TALiP dimensions (knowledge, purpose, design/implementation, interpretation/use).

Author(s)'Year	**Context/Participants**	**Methodology**	**Focus Area**	**Key Findings/Themes**
Rehmani (2003)	Secondary public exams (SSC/HSSC)	Case study	Effect of high-stakes exams on pedagogy, learning and assessment reforms.	Public examination system exerts strong negative washback, narrowing teaching to test preparation.
Ashraf & Zaki (2019)	Pakistani higher education; 30 English-language teachers from 17 Karachi universities (public & private).	Qualitative semi-structured questionnaires and in-depth interviews plus exam-paper content analysis	Challenges in language assessment	Low assessment literacy and institutional/practical constraints produce narrow, writing-heavy assessment with negative washback.
Gul et al., (2023)	Pakistani society and education system; focus on elite vs lower-class dynamics in English assessment.	Qualitative document analysis of Dawn newspaper reports and research texts (thematic/content analysis).	Politics, power, and socio-economic implications of English language assessment in Pakistan.	English assessment reinforces elitism, linguistic imperialism, and top-down policy control, shaping access to jobs, education and power.
Ilyas & Zamir, (2020)	University teachers in Pakistan	Qualitative study using literature review, HEC/HEI practices, and informal interviews with teachers guided	Continuous professional development (CPD) challenges and prospects in higher education	Resource gaps, weak policies, Sporadic workshops insufficient for sustained LAL development.
Hidri et al., 2024	735 English-language teachers across Pakistan (primary, secondary, vocational, university),	Cross-sectional survey using Brown's Teachers' Conceptions of Assessment (TCoA) abridged	How Pakistani English teachers conceptualize assessment	Assessment for accountability and improvement, but Brown's TCoA only partially fits Pakistan's exam-driven context,

		inventory		revealing policy–practice misalignment and a need for assessment-literacy PD
Shahzad, Hussain, and Habib (2022)	Two Pakistani universities, 134 English teachers	Mixed Methods	Formative assessment practices	Formative assessment is sporadic, weakly documented, and accuracy-centric, lacking standardized rubrics and communicative focus
Asif & Naz (2021)	Undergraduate ESL classes	Qualitative (N=12)	Assessment methods	Teachers rely on traditional written tests; formative assessment rarely used.
Husan & Shakur (2023)	Secondary teachers	Survey & focus groups	Perceptions	Teachers equate assessment with summative exams; formative seen as peripheral.
Fazli et al. (2024)	University EFL teachers	Survey (N=200)	Practices & LAL	Teachers report limited confidence in designing authentic assessment tasks.

2.7 Quality Appraisal

Quality was assessed using the Mixed Methods Appraisal Tool (MMAT 2018) for empirical studies and conceptual-coherence criteria for theoretical or policy-oriented works. Appraisal considered the clarity of research questions, appropriateness of study design, methodological rigor, transparency of analysis and relevance to the Pakistani context. Lower-quality studies were retained to avoid unnecessary exclusion from a relatively small evidence base, but their findings were interpreted cautiously and were not treated as equivalent in evidential strength to stronger studies. The review reports the appraisal procedure but does not contain the underlying item-level MMAT rating sheet. This reporting limitation is acknowledged explicitly rather than retrospectively estimating study quality. The review did not systematically conduct backward/forward citation snowballing or a separate structured search of theses and dissertations. These sources may contain additional Pakistani evidence and are therefore acknowledged as a coverage limitation. The search nevertheless included major international databases, Google Scholar and Pakistani journals/repositories within the defined eligibility criteria.

2.8 Synthesis Approach

Given heterogeneity in study designs and outcomes, a narrative thematic synthesis was employed. Studies were coded into thematic categories derived from TALiP and teacher cognition frameworks. Patterns were compared across studies to identify convergences, divergences and gaps. No statistical meta-analysis was possible due to the absence of comparable quantitative effect sizes.

3. Results

3.1 RQ1: Teachers' Conceptions and Purposes of Assessment

Across the evidence included, teachers were frequently described as equating assessment with summative testing, particularly high-stakes board and university examinations (Husan & Shakur, 2023; Hidri et al., 2024). Although some teachers could distinguish diagnostic, formative and summative purposes conceptually, enacted assessment was commonly organized around examination performance. This distinction is analytically important: knowledge of assessment terminology did not necessarily translate into assessment-for-learning practices. The pattern is consistent with the teacher-cognition perspective in which contextual expectations and prior experience shape which forms of assessment teachers regard as legitimate or practicable.

Teachers' rationales reflected contextual pressures: student futures depended on exam results which reinforced "assessment *of* learning" over "assessment *for* learning." Very few studies reported strong integration of formative or diagnostic assessment cycles.

3.2 RQ2: Classroom Practices and Tools

Across studies addressing classroom practice, assessment was dominated by written tests, multiple-choice items and examination-oriented essays (Asif & Naz, 2021; Fazli et al., 2024; Mohammad & Javed, 2019; Shah & Afzal, 2019). Alternative approaches, including portfolios, peer review and project-based tasks, appeared less frequently and were generally associated with particular institutional or program contexts (Shahzad et al., 2022). Speaking and listening assessment was especially constrained where external examinations did not require these skills and where trained raters, time or equipment were unavailable. The convergence across these studies suggests that practicality is not merely a technical

consideration; in the Pakistani context it interacts with examination demands to narrow the construct actually assessed.

Speaking and listening assessments were almost absent in mainstream classrooms because they were not part of external exams and required resources (e.g., audio equipment, trained raters) that schools typically lacked. Writing assessment was reported but often marked inconsistently with rubrics either absent or unused (Shahzad et al., 2022).

In short, teachers prioritized practicality and reliability (easy-to-mark formats) over authenticity and validity.

3.3 RQ3: Interpretation and Use of Assessment Evidence

Studies addressing interpretation and use of assessment evidence generally indicated that results were communicated as grades, scores or corrected scripts, with limited evidence that assessment information was systematically used to modify subsequent instruction (Shahzad et al., 2022; Ahmad & Akbar, 2020). Where formative feedback occurred, it was often teacher-initiated rather than embedded in institutional assessment cycles. This finding matters because assessment literacy involves not only producing scores but also interpreting evidence and using it to make pedagogical decisions. The reported gap therefore concerns the full assessment process rather than test construction alone.

Where feedback cycles existed, they were idiosyncratic and teacher-driven rather than institutionalized practices. Concerns around fairness such as cultural bias in texts or unequal access to tuition were occasionally noted but rarely systematically addressed (McNamara, 2000; Shohamy, 2001).

3.4 RQ4: Teacher Education and Professional Development in LAL

The evidence on teacher preparation indicates that pre-service education often provides general exposure to assessment but limited language-specific preparation in areas such as rubric development, performance assessment and rating productive skills (Ilyas & Zamir, 2020; Husan & Shakur, 2023). In-service development was frequently described as episodic rather than sustained, with workshops offering limited opportunities for teachers to trial, evaluate and refine assessment practices. This helps explain why awareness of alternative assessment principles can coexist with continued reliance on familiar examination formats.

As a result, teacher confidence in using diverse assessment methods remained low. Most educators reverted to inherited exam formats perceived as safe and legitimate. Scarino's (2013) argument for assessment literacy as a reflective practice had little presence in the Pakistani evidence base.

3.5 RQ5: Systemic Challenges (Washback, Practicality, Fairness)

Negative washback from high-stakes examinations was the most recurrent systemic theme. Studies linked examination requirements with rote memorization, curriculum narrowing and reduced attention to oral/aural skills (Rehmani, 2003; Aftab et al., 2014; Sajjad & Bughio, 2018). Practical constraints including large classes, marking workload and limited resources further encouraged easily administered written formats (Fazli et al., 2024). Fairness concerns were also connected to unequal access to private tuition and other forms of examination preparation (Gul et al., 2023). These findings indicate that teachers' assessment choices cannot be interpreted solely as individual deficits in LAL; they are embedded in a system in which examination outcomes carry high academic and social consequences.

Practicality issues like large classes, heavy marking loads and lack of resources further reinforced reliance on traditional written exams (Fazli et al., 2024). Fairness concerns included socio-economic inequalities, as wealthier students accessed private tuition or test-prep resources, widening achievement gaps (Shohamy, 2001; Gul et al., 2023).

Some innovations were reported (e.g., portfolio trials, project-based assessments), but these remained micro-level reforms confined to better-resourced institutions (Asif & Naz, 2021).

3.6 Study Characteristics Overview

As shown in Table 1, the included studies span from 2003 to 2024 covering secondary, college and university contexts. Most were conducted in Punjab and urban centres with relatively fewer focusing on rural or under-resourced areas. Methodologies included surveys, interviews, mixed methods and policy analyses. Overall, evidence demonstrates strong consistency across levels: exam-centric culture, weak formative practices, underdeveloped LAL and systemic washback dominate the findings.

4. Discussion and Implications

The synthesis identifies a coherent pattern across the five research questions: Pakistani English language teachers' assessment practices are predominantly examination-oriented, formative use of assessment evidence is comparatively weak, professional preparation in LAL is uneven and systemic conditions especially high-stakes examinations, workload and resource constraints shape what teachers can realistically implement. The central interpretive point is that these patterns should not be reduced to a simple deficit model of teachers. The evidence instead indicates an interaction between teacher knowledge and cognition and the institutional ecology in which assessment decisions are made. This interpretation is important because professional development alone is unlikely to produce durable change if examination incentives continue to reward narrow forms of assessment.

4.1 Interpreting the Findings through TALiP and Teacher Cognition

Viewed through TALiP, the findings suggest uneven development across the assessment process. Teachers appear more familiar with established purposes and item formats than with designing authentic performance tasks, constructing usable rubrics, rating productive skills or systematically interpreting evidence for subsequent instruction (Shahzad et al., 2022; Fazli et al., 2024). The distinction between knowing about assessment and enacting assessment is therefore central. Teacher cognition theory helps explain this gap: beliefs about what counts as legitimate assessment are formed within institutional and examination environments, and practices that are efficient, familiar and aligned with external tests can remain attractive even when teachers recognize the pedagogical value of alternatives. The evidence consequently supports a practice-in-context interpretation of LAL rather than a purely individual competence model.

The findings also complicate the assumption that examination-oriented practice is simply resistant to reform. From the teacher's perspective, conventional formats may represent rational adaptations to large classes, heavy marking loads, limited resources and consequential examinations. In such circumstances, a teacher who substitutes an unfamiliar performance assessment may incur additional workload without receiving institutional recognition or examination benefit. The persistence of traditional practice is therefore partly an incentive problem. This explains why isolated training interventions may increase teachers'

knowledge without substantially changing classroom assessment. Reform must alter both teachers' assessment repertoires and the conditions under which those repertoires are enacted.

4.2 Positioning Pakistan in the Global LAL Literature

The Pakistani evidence broadly reflects international LAL research showing that teachers often experience difficulties in formative assessment and assessment-for-learning practices (Fulcher, 2012; Vogt & Tsagari, 2014; Tsagari & Vogt, 2017). What is distinctive in the present synthesis is the extent to which high-stakes examination structures appear to mediate these difficulties. The Pakistani context therefore illustrates a broader principle in LAL research: assessment literacy is situated. Knowledge of validity, fairness or formative assessment does not automatically determine practice when external assessments communicate stronger incentives. The synthesis consequently supports context-sensitive LAL models that incorporate institutional, policy and examination conditions alongside teacher knowledge and beliefs.

4.3 Implications for Policy Makers and Exam Boards

The policy implication is not that high-stakes examinations should simply be removed, which would be unrealistic and could introduce new equity and comparability problems. Rather, the construct represented in high-stakes examinations should be progressively broadened so that important curricular competencies are also rewarded. Manageable speaking and listening components, carefully piloted performance tasks, transparent specifications and moderation procedures could gradually reduce the mismatch between communicative curricula and written examination demands. At the same time, reforms should be evaluated for fairness and reliability because poorly designed performance assessment could reproduce rather than solve inequities. The evidence therefore supports phased alignment rather than abrupt replacement of existing systems.

- ***Broaden constructs****:* Incorporate manageable speaking and listening components (e.g., recorded prompts, paired oral tasks) into national exams to incentivize classroom attention to communicative skills.
- ***Publish specifications****:* Provide detailed exam blueprints, rubrics and marked exemplars to guide teachers away from speculation and drilling.

- ***Pilot innovations****:* Initiate phased reforms in selected districts, combining performance tasks with statistical moderation to safeguard fairness and reliability.
- ***Ensure fairness****:* Commission fairness audits to examine cultural bias in exam texts and disparities in access to resources.

Such steps would not only reduce negative washback but also elevate the construct validity of assessments, aligning them more closely with curricular goals of communicative competence.

4.4 Implications for Teacher Education and Professional Development

The professional-development implications follow directly from the practice-in-context interpretation. The evidence does not suggest that another generic workshop on assessment terminology would be sufficient. Teacher education should address concrete assessment practices task design, rubric construction, rating, feedback and interpretation of evidence and should provide opportunities for teachers to apply these practices to their own assessment materials. Coaching, peer moderation and communities of practice are potentially more appropriate than isolated workshops because they connect assessment knowledge with classroom enactment. Such provision should also be designed around the constraints identified in the review, including large classes and limited resources, rather than importing assessment models that require infrastructure unavailable to many institutions.

Teacher education programs must move beyond cursory treatment of assessment concepts. To build sustainable language assessment literacy, pre-service and in-service programs should:

- ***Embed language-specific training****:* Include modules on designing tasks for speaking, listening, and writing, developing analytic rubrics, and conducting moderation.
- ***Shift from workshops to coaching****:* Replace one-off training sessions with iterative cycles where teachers design tasks, trial them in classrooms and receive feedback on student work.
- ***Promote reflective practice****:* Encourage teachers to analyse their own assessment artifacts (tests, rubrics, graded scripts), fostering the self-awareness emphasized by Scarino (2013).

- ***Develop communities of practice****:* Establish departmental moderation groups or peer observation schemes to normalize discussion of assessment quality and fairness.

These approaches resonate with global calls for learning-oriented assessment (Carless, 2015) but must be adapted to Pakistan's large classes and resource constraints.

4.5 Implications for School and College Leaders

Institutional leadership is a further mediating level. Teachers may possess assessment knowledge but still avoid alternative practices if departments do not provide time for feedback, moderation or collaborative planning. School and college leaders can therefore influence LAL indirectly by protecting time for assessment cycles, supporting peer review of assessment materials and monitoring the balance between summative and formative tasks. These measures are relatively low-cost compared with examination reform and could provide an intermediate route through which policy intentions become classroom routines.

At the institutional level, principals and department heads have a crucial role in enabling assessment reform. They can:

- ***Protect instructional time****:* Allocate time for formative cycles (e.g., "feedback weeks" or re-teach sessions).
- ***Reward innovation****:* Recognize and incentivize teachers who pilot alternative assessments and systematically use evidence to inform teaching.
- ***Audit assessment balance****:* Review departmental assessment practices across a term to ensure that formative and authentic tasks complement traditional tests.

Leadership support can mitigate the risks teachers perceive in adopting non-traditional methods.

4.6 Implications for Researchers

The evidence base itself reveals an important methodological gap. Many studies rely on teacher self-report, while fewer directly examine assessment artefacts, observed classroom assessment episodes or student experiences. This limits confidence in the relationship between reported beliefs and practice enacted. The geographical concentration of studies in Punjab and urban settings further limits the transferability of conclusions to rural and under-

resourced contexts. Future studies should therefore combine interviews and surveys with observation, assessment-document analysis, and student evidence and should explicitly compare institutional contexts rather than treating Pakistan as a homogeneous assessment environment.

This review also highlights gaps in the evidence base. Most studies are clustered in Punjab and urban areas; rural and under-resourced contexts remain underrepresented. Moreover, reliance on self-reports limits insight into actual classroom practices. Future research should:

- Conduct classroom observations and artifact analyses (e.g., teachers' tests, rubrics) to triangulate reported practices.
- Explore student perspectives, especially regarding fairness and feedback.
- Examine equity dimensions, such as socio-economic disparities in access to assessment preparation.
- Pilot and evaluate interventions (e.g., rubric training, portfolio assessment) to test scalable models of LAL development.

4.7 Synthesis of Implications

Taken together, the findings support a two-level theory of change. At the teacher level, LAL development should strengthen practical knowledge, judgment and confidence in using assessment for learning. At the system level, examination specifications, institutional expectations and resource arrangements should provide incentives for those practices to be used. The two levels are mutually dependent: teacher development without systemic alignment is likely to remain localized, while examination reform without teacher capacity may produce compliance rather than meaningful change. This dual-alignment argument is the principal interpretive contribution of the synthesis.

5. Limitations

Several limitations should be considered when interpreting the synthesis. First, the evidence base is geographically uneven, with many studies conducted in Punjab and urban centers; rural and under-resourced settings remain less visible. Second, many included studies rely on self-reported perceptions, with fewer using classroom observations, assessment artefacts or student outcome data. Third, the review was restricted to English-language publications and

did not include a systematic search of theses, dissertations or other grey literature. Backward and forward citation snowballing was also not conducted. These choices may have excluded relevant evidence. Fourth, heterogeneity in study designs and outcomes precluded statistical meta-analysis, so findings were synthesized narratively. Fifth, although MMAT and conceptual-coherence criteria were applied, lower-quality studies were retained and interpreted cautiously. Finally, the screening process involved two reviewers, but no formal inter-rater agreement statistic was calculated; disagreements were resolved through discussion. This procedural deviation should be considered when assessing the reproducibility of the selection process.

Third, the review was restricted to English-language publications. Grey literature, theses or Urdu-language publications may contain additional insights that were excluded. Fourth, the heterogeneity of study designs precluded any meta-analysis, and thus findings were synthesized narratively. Fifth, while a quality appraisal (using MMAT and conceptual coherence criteria) was conducted, lower-quality studies were retained for inclusivity, which may introduce bias. Finally, the review protocol was not pre-registered, which may affect transparency compared to reviews with registered protocols.

Despite these limitations, the review consolidates a substantial body of evidence across three decades and provides a structured synthesis of teachers' assessment perceptions, practices, professional preparation and systemic constraints in Pakistan. The contribution should therefore be understood as evidence synthesis rather than a nationally representative estimate of teacher behaviour.

1. Conclusion

This systematic review has brought together three decades of research to answer five interrelated questions about Pakistani English language teachers' conceptions, practices, use of evidence, professional development, and systemic challenges in language assessment. The findings paint a consistent picture: teachers perceive assessment primarily as summative testing, classroom practices mirror board examination formats, interpretation of assessment evidence is limited, professional development in LAL remains fragmented, and systemic washback from high-stakes exams constrains innovation.

6.1 Answering the Research Questions

- **RQ1 (Conceptions):** Teachers largely conceptualize assessment as a tool for certification and accountability rather than for diagnosis or learning.
- **RQ2 (Practices):** Written, exam-style tests dominate classroom practice while authentic or performance-based assessments remain rare.
- **RQ3 (Interpretation/Use):** Feedback is typically reduced to grades and corrected answers, with little evidence of formative use to adjust teaching.
- **RQ4 (Professional Development):** Pre-service and in-service training in assessment is limited in scope and depth, leaving teachers ill-prepared to implement diverse strategies.
- **RQ5 (Systemic Challenges):** Negative washback, large class sizes and fairness concerns are pervasive, reinforcing conservative assessment practices.

6.2 Theoretical Contributions

A central contribution of this review is that, to our knowledge, it provides the first PRISMA-guided systematic synthesis of research on English language teachers' assessment perceptions, practices and challenges in Pakistan. Rather than treating individual studies as isolated accounts, the review brings together evidence across educational levels and examines recurring patterns through TALiP and teacher cognition theory. The synthesis shows that the assessment problem is not simply one of insufficient teacher knowledge but of practice-in-context: teachers' assessment decisions are mediated by examination structures, institutional expectations, workload, resources and established beliefs.

6.3 Practical and Policy Contributions

By integrating the TALiP framework with teacher cognition theory, the review extends interpretation beyond what teachers know and do to why particular practices persist. Teachers' choices can be understood as rational responses to a high-stakes ecology in which examination outcomes carry substantial consequences. This suggests that LAL development should be conceptualized as both an individual professional capacity and a situated practice shaped by systemic incentives.

The review highlights several actionable reforms. For teachers, practical routines such as simplified rubrics, brief oral checks, or staged writing tasks can foster formative cycles even

in large classes. For teacher educators, the shift from workshop-driven PD to coaching and communities of practice is crucial for sustained LAL development. For exam boards and policymakers, aligning assessment formats with curricular competencies by gradually introducing communicative tasks and ensuring fairness audits can reduce negative washback and incentivize innovation.

6.4 Implications for Global LAL Discourse

Pakistan's case illustrates how structural constraints can amplify familiar challenges in language assessment literacy. While global research identifies persistent gaps in formative assessment use across contexts, the exam-driven culture in Pakistan makes these gaps particularly resistant to change. The synthesis therefore contributes to global LAL discourse by demonstrating how assessment literacy is shaped not only by teachers' knowledge and beliefs but also by examination regimes, institutional resources and policy expectations. It reinforces the need for context-sensitive models of LAL that recognize practical constraints while advancing validity, fairness and communicative competence.

6.5 Future Directions

Future research should expand coverage to under-researched regions, employ triangulated methods (observations, artifact analysis, student data) and test scalable interventions such as rubric-based writing assessment or moderated portfolios. Policymakers should institutionalize partnerships between universities, schools, and exam boards to co-design reforms. Without systemic support, classroom innovations will remain isolated and fragile.

6.6 Final Reflection

Ultimately, this review underscores a paradox: teachers in Pakistan recognize the value of assessment, yet their practices remain constrained by exams, workloads, and limited training. Breaking this cycle requires dual alignment, empowering teachers with the tools and confidence to use assessment for learning, and re-engineering high-stakes examinations to reward communicative competence, fairness and authenticity. Only when these two levels converge will classroom assessment in Pakistan begin to reflect the broader educational goals of equity, quality, and meaningful learning.

References

Abbasi, S. (2022). *A Comparative Analysis of Institutional Practices for The Evaluation in Teachers' Education Programs*. *1*(2), 26–42. https://doi.org/10.55464/pjar.v1i2.17

Aftab, A., Qureshi, S., & William, I. (2014). Investigating the washback effect of the Pakistani Intermediate English Examination. *International Journal of English and Literature, 5*(7), 149–154. https://doi.org/10.5897/IJEL2014.0595

Ahmad, I., & Akbar, R. A. (2020). Examining the relationship between self-efficacy beliefs of elementary level English teachers and their implementation practices of formative assessment in Punjab. *Review of Education, Administration & Law, 3*(2), 123–134.

Ajmal, A., Khalid, A., & Nazir Ahmad, S. (2022). Classroom assessment practices of English teachers at secondary school level in Punjab, Pakistan. *Journal of Educational Research, 25*(1), 45–60.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115–129. https://doi.org/10.1093/applin/14.2.115

Ali, I. (2021). Exploring how ELT teachers perceive and practice English language assessment. *Journal of Language and Education*, *7*(1 (25)), 109-126.

Ali, A., & Saeed, M. (2024). Teachers' Perceptions and Practices about Grade V and VIII Students' Assessment under Punjab Examination Commission. *Pakistan Social Sciences Review*, *8*(3), 448-462.

Ashraf, A., & Zaki, S. (2019). Language assessment in higher education: Challenges and consequences. *Pakistan Journal of Education, 36*(3), 163–185.

Asif, R., & Naz, S. (2021). Analysis of Assessment Methods Used by ESL Teachers at Undergraduate level in Pakistan. *Global Language Review*, *6*(2), 191-200.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2001). Designing and developing useful language tests. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 109–116). Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Bond, M. (2021). Schools and emergency remote education during the COVID-19 pandemic: A living rapid systematic review. *Asian Journal of Distance Education, 15*(2), 191–247.

Borg, S. (2003). Teacher cognition in language teaching: A review of research on what language teachers think, know, believe, and do. *Language Teaching, 36*(2), 81–109. https://doi.org/10.1017/S0261444803001903

Borg, S. (2015). *Teacher cognition and language education: Research and practice*. Bloomsbury.

Broadfoot, P., Daugherty, R., Gardner, J., Harlen, W., James, M., & Stobart, G. (2002). Assessment for learning: 10 principles. *Assessment Reform Group*.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Pearson.

Brown, H. D. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). Pearson.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge University Press.

Carless, D. (2007). Learning-oriented assessment: Conceptual bases and practical implications. *Innovations in Education and Teaching International, 44*(1), 57–66. https://doi.org/10.1080/14703290601081332

Carless, D. (2015). *Excellence in university assessment: Learning from award-winning practice*. Routledge.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge University Press.

Colby-Kelly, C., & Turner, C. E. (2007). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *Canadian Modern Language Review, 64*(1), 9–37.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25*(3), 327–347. https://doi.org/10.1177/0265532208090156

Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly, 43*(3), 393–415.

Davis, J. M., Norris, M. E., Malone, T. H., McKay, T. H., & Son, Y.-A. (Eds.). (2018). *Useful assessment and evaluation in language education*. Georgetown University Press.

Dewey, J. (1960). *On experience, nature and freedom: Representative selections* (R. J. Bernstein, Ed.). Bobbs-Merrill.

Dewey, J. (1973). *The philosophy of John Dewey*. Putnam's Sons.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.

Douglas, B. (2016). *Language assessment: Principles and classroom practices*. Pearson.

Eckes, T. (2023). *Introduction to many-facet Rasch measurement*. Peter Lang.

Fazli, S. K., Azim, M. U., & Saeed, M. I. (2024). Language assessment practices of in-service English language teachers at Pakistani universities. *Journal of Policy Research, 10*(2), 768–775. https://doi.org/10.61506/02.00296

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132. https://doi.org/10.1080/15434303.2012.703635

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. Routledge.

Giraldo, F. (2018). Language assessment literacy: Implications for language teachers. *Profile: Issues in Teachers' Professional Development, 20*(1), 179–195. https://doi.org/10.15446/profile.v20n1.64664

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education*. Cambridge University Press.

Green, A. (2013). *Exploring language assessment and testing: Language in action*. Routledge.

Gul, N., Ali, A., Hassan, S. S. U., & Rasheed, A. (2023). An investigation into the politics of English language assessment in Pakistan. *Journal of Policy Research, 9*(4), 84–93.

Husan, S., & Shakur, N. (2023). TEACHERS'PROFICIENCY IN ENGLISH LANGUAGE ASSESSMENT AT AN ENGLISH-MEDIUM UNIVERSITY: IMPLICATIONS FOR ELT TRAINING IN PAKISTAN. *Journal of Social Sciences Development*, *2*(2), 339-354.

Husain, N. (2013). *ESL teachers' perceptions and practices on formative assessments in classroom*. http://eprints.utm.my/id/eprint/38014/

Hussain, S., Shaheen, N., Ahmad, N., & Islam, S. U. (2019). Teachers' classroom assessment practices: challenges and opportunities to classroom teachers in Pakistan. *Dialogue*, *14*(1), 88-97.

Ilyas, M., & Zamir, S. (2020). *Cpd of university teachers in pakistan: challenges and prospects*. *1*, 1–12. https://issrapapers.ndu.edu.pk/site/article/view/55

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*(3), 385–402. https://doi.org/10.1177/0265532208090158

Inbar-Lourie, O. (2017). Language assessment literacy. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (pp. 257–270). Springer. https://doi.org/10.1007/978-3-319-02261-1_19

Khan, E. A., & Hussain, B. (2019). Grade Inflation in Universities of Pakistan: A Case of Bahauddin Zakariya University, Multan. *Pakistan Journal of Social Sciences (PJSS)*, *39*(3).

Khan, S. D., & Mohammed, L. A. (2024). Assessing the Impact of Student's Evaluation of Teachers (SET) on English Speaking and Writing Anxiety among BS English Students in Peshawar, Pakistan. *International Journal of Emerging Issues in Social Science, Arts and Humanities (IJEISSAH)*, *2*(3), 85-93.

Leung, C. (2005). Convivial communication: Recontextualizing communicative competence. *International journal of applied linguistics*, *15*(2), 119-144.

Leung, C. (2014). Classroom-based assessment issues for language teacher education. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 3, pp. 1515–1529). Wiley. https://doi.org/10.1002/9781118411360.wbcla083

Leung, C., & Scott, C. (2009). Formative assessment in language education policies: Emerging lessons from Wales. *Annual Review of Applied Linguistics, 29*, 64–79. https://doi.org/10.1017/S0267190509090060

Leung, C., & Scott, C. (2010). English language teaching and teacher education in Hong Kong. *Language Teaching, 43*(3), 249–264.

Leung, C., & Taylor, B. (2013). Assessment quality and standards: The role of language teacher associations. *Language Assessment Quarterly, 10*(2), 247–267. https://doi.org/10.1080/15434303.2013.769547

McNamara, T. (2000). *Language testing*. Oxford University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.

Mickan, P. (2017). *Text-based research and teaching: A social semiotic perspective on language in use*. Springer.

Mohammad, R., & Javed, S. (2019). Assessment practices of English teachers at college level in Pakistan. *Pakistan Journal of Social Sciences, 39*(4), 1461–1472.

Naveed, S., & Anwar, S. (2020). Challenges of implementing classroom assessment in Pakistani schools. *Bulletin of Education and Research, 42*(1), 55–70.

O'Sullivan, B. (2012). Assessment issues in language testing. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–6). Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0053

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice, 48*(1), 4–11. https://doi.org/10.1080/00405840802577536

Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal, 100*(S1), 190–208. https://doi.org/10.1111/modl.12308

Rahman, T. (2002). *Language, ideology and power: Language-learning among the Muslims of Pakistan and North India*. Oxford University Press.

Rehmani, A. (2003). Impact of public examination system on teaching and learning in Pakistan. *International Biannual Newsletter ANTRIEP*, *8*(2), 3-7.

Rea-Dickins, P. (2008). Classroom-based language assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7, pp. 257–271). Springer. https://doi.org/10.1007/978-0-387-30424-3_153

Saeed, M., & Mahmood, T. (2020). Language assessment practices at higher secondary level in Punjab. *Pakistan Journal of Education, 37*(2), 89–106.

Sajjad, M., & Bughio, F. A. (2018). Washback effect of English language testing on teaching and learning. *International Journal of English Linguistics, 8*(4), 116–122. https://doi.org/10.5539/ijel.v8n4p116

Shah, S. R., & Afzal, H. (2019). Teachers' assessment practices at elementary level in Punjab. *Bulletin of Education and Research, 41*(1), 87–100.

Shahzad, K., Hussain, B., & Habib, A. (2022). Analyzing formative assessment practices of english language teachers in pakistan. *Pakistan Journal of Social Research*, *04*(01), 61–70. https://doi.org/10.52567/pjsr.v4i1.345

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. Pearson Education.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*(7), 534–539.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238–245.

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice, 18*(1), 23–27. https://doi.org/10.1111/j.1745-3992.1999.tb00004.x

Stiggins, R. J. (2010). Essential formative assessment competencies for teachers and school leaders. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 233–250). Routledge.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403–412. https://doi.org/10.1177/0265532213480338

Taylor, L. (Ed.). (2014). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge University Press.

Tsagari, D., & Vogt, K. (2017). Assessment literacy of foreign language teachers around Europe: Research, challenges and future prospects. *Papers in Language Testing and Assessment, 6*(1), 41–63.

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374–402. https://doi.org/10.1080/15434303.2014.960046

Wiliam, D. (2011). *Embedded formative assessment*. Solution Tree Press.

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58*, 149–162. https://doi.org/10.1016/j.tate.2016.05.010

Zafar, S., & Mehmood, A. (2018). Teachers' perceptions and practices of formative assessment in ESL classrooms. *Bulletin of Education and Research, 40*(1), 79–90.