

Reliability and Validity of Locally Developed English Proficiency Tests: A Case Study of the Department of English, Sulugh Campus, University of Benghazi, Libya

Farag Malik Alfakhiraa

Department English Faculty Arts and Science Sulugh, University Benghazi, Sulygh, Libya

farg.m.elfakhri@uob.edu.ly

ثبات وصدق اختبارات كفاءة اللغة الإنجليزية المطورة محلياً: دراسة حالة لقسم اللغة الإنجليزية فرع سلوك جامعة بنغازي

فرج مالك الفاخري

قسم اللغة الإنجليزية، كلية الآداب والعلوم سلوك، جامعة بنغازي، سلوك، ليبيا

Received: 08/9/2026; Accepted: 2026/9/10; Published 15/9/2026

Abstract

This case study explores the reliability and validity of locally developed English proficiency tests administered in the Department of English, Sulugh Campus, University of Benghazi, Libya. The study adopted a convergent mixed-methods approach combining a quantitative analysis of test scores from two academic cohorts (2021/2022 and 2022/2023) with qualitative data from semi-structured interviews with five instructors, two focus groups with twenty undergraduate students, and ten non-participant classroom observations. Internal consistency, measured using Cronbach's alpha, was .65 to .75 across cohorts ($\alpha = .72$ and $\alpha = .68$, respectively), reflecting moderate reliability, which is sufficient for group-level interpretation, but not high-stakes decisions about individual students. However, qualitative findings indicate considerable validity issues: around 60-70% of total marks were said to comprise discrete-point grammar and vocabulary items, and speaking was virtually excluded from formal assessment. Instructors and students agreed that the tests were narrowly focused on isolated linguistic knowledge, rather than on communicative competence, and were poorly aligned with descriptors of the Common European Framework of Reference for Languages (CEFR). The classroom observations also showed a widespread negative washback, since the instruction focused on test preparation rather than overall language development. The study concludes that the tests are sufficiently reliable, but their limited content and construct validity hinder their ability to measure real-world communicative ability. The study recommends re-designing test blueprints around authentic communicative tasks, providing structured assessment-literacy training for faculty, and embedding formative assessment to reduce negative washback and better serve Libyan EFL learners.

Keywords: Test reliability, test validity, communicative competence, washback effects, Libyan EFL learners

ملخص البحث

تفحص هذه الدراسة حالة موثوقية وصلاحيّة اختبارات الكفاءة في اللغة الإنجليزية المصمّمة محلياً في قسم اللغة الإنجليزية، حرم سلوك، جامعة بنغازي، باستخدام تصميم مختلط من المنهجين. جمعت الدراسة بين تحليل كمي لدرجات الاختبارات لدفعتين دراسيتين (2021/2022 و 2022/2023)، وبيانات نوعية جمعت من مقابلات شبه منظمّة مع خمسة من المُدرّسين، ومجموعتي تركيز ضمنا عشرين طالباً، وعشر مشاهدات صفيّة غير مشاركة. أظهرت قيمة ألفا كرونباخ اتساقاً داخلياً متوسطاً تراوح بين 0.65 و 0.75، بما يدعم موثوقية الاختبارات على المستوى الجماعي، دون القرارات ذات المخاطر العالية المتعلّقة بالطلاب الأفراد. ومع ذلك، كشفت النتائج النوعيّة عن قصور جوهري في الصلاحيّة، إذ تستأثر بنود القواعد والمفردات المنفصلة نحو 60 - 70% من إجمالي الدرجات، بينما تغيب مهارة التحدّث شبه الكامل عن التقويم الرسمي. وقد أجمع المُدرّسون والطلاب على أنّ الاختبارات لا

تتوافق مع الأهداف التواصلية أو مع معايير الإطار الأوروبي المرجعي المشترك. وأظهرت المشاهدات الصقيّة أنّ التدريس يميل إلى الدوران حول التحضير للاختبار بدلاً من تنمية اللغة على نحوٍ أشمل. وتخلّص الدراسة إلى أنّه رغم الموثوقية المقبولة، تبقى الصلاحية المحتوائية والبنائية المحدودة للاختبارات تقيّد قدرتها على قياس القدرة التواصلية الفعلية. وتوصي الدراسة بإعادة صياغة مخطّطات الاختبارات حول مهامّ تواصلية أصيلة، وتوفير تدريبٍ مُوجّه في الثقافة التقويمية لأعضاء هيئة التدريس، ودمج أساليب التقويم التكويني للحدّ من الآثار الارتدادية السلبية.

الكلمات المفتاحية: موثوقية الاختبار، صلاحية الاختبار، الكفاءة التواصلية، الآثار الارتدادية، متعلمو اللغة الإنجليزية كلغة أجنبية في ليبيا

1. Introduction

Valid and reliable language testing is not a peripheral technicality; it is the bedrock of effective teaching, sound curriculum development and fair learner outcomes. The stakes are especially high in Libya, where access to higher education, employment, and international mobility increasingly depends on demonstrated English proficiency. However, there are still concerns about whether locally developed tests can provide an accurate picture of learners' actual language ability.

In recent years, English proficiency has become increasingly important in Libya, with universities, government ministries and private-sector employers requiring internationally recognisable standards of communicative competence for scholarships, employment and academic mobility. Unfortunately, there is little formal, standardised testing infrastructure at the departmental level that compares to internationally recognised exams. In practice, this has meant that individual departments, such as the Department of English at Sulugh Campus, are left to design, administer and score their own proficiency tests with little external oversight, standardisation, or psychometric validation.

Reliability is the consistency of scores across items, raters, and administrations. A test has to be valid, i.e. it measures what it is supposed to measure, in this case the ability to use English communicatively, not the rote learning of grammar rules and vocabulary lists. The practice of locally developing tests is common in the majority of Libyan university settings including the Department of English, Sulugh Campus, University of Benghazi. It rarely meets the psychometric standards expected of modern language assessments.

This decentralisation of test development is not unique to the Sulugh Campus. Across the wider Libyan higher education system, locally-developed examinations are typically produced by individual instructors or small departmental committees, usually without formal training in test design, item analysis or validation procedures. Given the lack of standardised quality-assurance mechanisms, the psychometric properties of these tests, i.e., their reliability and, in particular, their validity, are rarely empirically examined, regardless of the fact that the tests have far-reaching consequences for students' academic progression, certification and eventual employability.

Recent empirical work has provided greater justification for perceiving locally developed language tests as evidence-based systems of measurement rather than as routine classroom examinations. Noroozi and Karami (2024) demonstrated via a Rasch validation of a university English proficiency test that robust internal reliability can be accompanied by less conclusive evidence for certain aspects of validity. Sureeyatanapas et al.(2024) further emphasized the need for inter-rater and intra-rater reliability for English performance-based assessments as scoring inconsistency may threaten the fairness and accuracy of proficiency judgments.

Quality of assessment is also affected by the testing culture of the institution and assessment knowledge of the teachers. A recent systematic review found that EFL teachers commonly require further professional development in language assessment literacy and assessment-related practice

(Puspawati et al., 2024), while classroom-based assessment research has highlighted the interaction between instructors' assessment approaches and the broader assessment culture in which they work (Latif & Wasim, 2024). Besides, research on washback on a large scale shows that high-stakes English tests may have a prominent impact on the learners' outcomes, and the effects are different depending on the educational context (Zhang, 2024).

Recent evidence suggests that this is not an isolated problem. A document analysis of a high-stakes public EFL examination in Libyan schools revealed a continuing incongruence between stated aims of the examination and its actual content, with writing assessed only indirectly and speaking omitted altogether (Onaiba, 2024). Similar patterns of examination-driven, discrete-point testing have been linked to negative washback across multiple EFL contexts, narrowing instruction toward test preparation rather than the communicative competencies that curricula formally espouse (Rahman, et al., 2023).

There has been an increase in Libyan studies on curriculum design, teacher beliefs and classroom pedagogy (see Section 5) but few have triangulated quantitative reliability data with qualitative evidence on validity and washback within a single Libyan institution. The present study attempts to fill this gap by combining statistical reliability analysis with interviews of instructors, focus groups with students, and observations of classrooms, so as to provide a more rounded empirical account of the operation of locally-developed tests in practice, not just whether they produce consistent scores but what those scores really measure and how they affect the day-to-day behaviour of the classroom. This case study attempts to respond to these concerns by looking at the assessment practices in the Department of English in Sulugh Campus in an attempt to find out the strengths, gaps, and opportunities for improvement in its locally developed tests. If reliability and validity are compromised the wider educational system is at risk of losing credibility. A systematic re-evaluation is therefore timely and necessary.

The rest of the paper is structured as follows. Section 2 summarises the guiding research questions of the study and Section 3 describes the problem statement that motivates the question. Research purposes are presented in Section 4. Section 5 discusses relevant literature on Libyan EFL testing, washback, and assessment literacy. Section 6 describes the mixed methods design and procedures of the study. Section 7 presents the quantitative and qualitative results with support from summary tables. Sections 8-11 discuss findings, practical implications and recommendations followed by the study limitations and conclusion and a summary of the study's contribution to language testing practice in Libya.

2. Research Questions

This study is guided by three core questions:

1. What is the reliability of the locally developed English proficiency tests administered in the Department of English at Sulugh Campus?
2. To what extent do these tests reflect students' real-world English ability — particularly communicative competence — rather than rote linguistic knowledge?
3. What are teachers' and students' attitudes toward these tests, and what washback effects do the tests have on classroom teaching and learning?

The study seeks, through these questions, to illuminate broader issues of language testing in Libyan higher education and to lay an evidence-based foundation for improving the quality of assessment at Sulugh Campus and other similar institutions. .

3. Problem Statement

The precise measurement of English proficiency is crucial to the academic progress and international employment of Libyan students. Yet locally developed assessment instruments at Sulugh Campus, University of Benghazi seem to be operating variably in relation to this goal. These tests over-emphasize discrete knowledge of grammar and vocabulary at the expense of students' ability to use English communicatively under real-world conditions.

This imbalance does not fit with modern standards which value integrated, communicative competence, and is likely exacerbated by limited instructor training in modern assessment methods. Therefore, standardised testing practices are at risk of distorting the educational process itself, rewarding rote memorisation rather than authentic language use and ultimately misrepresenting students' real capabilities and limiting the skills they need to succeed academically and professionally in international settings.

4. Research Objectives

The main objective of this study is to critically analyse the local English language testing practices in the Department of English, Sulugh Campus, University of Benghazi. Specific objectives are:

- Reliability: assess the internal consistency of locally administered English proficiency tests, to see if they produce stable scores across items and across administrations.
- Validity: discuss construct validity (how much the test is related to communicative language proficiency) and content validity (how much the tests measure what they are supposed to measure).
- Perceptions: use qualitative data to examine teachers' and students' views on the assessments, including fairness, relevance, difficulty and effectiveness.
- Washback: Evaluate the impact of test format and content on curriculum priorities, classroom practice and student study behaviour and identify positive and negative washback effects.
-

5. Literature Review

Research on EFL teaching, learning, and assessment in Libya has expanded in recent years, driven by growing attention to the linguistic needs of students and to the influence of testing on classroom practice. This section reviews that literature under four themes: language testing in the Libyan context, washback effects on teaching and learning, teacher assessment literacy, and the theoretical grounding of reliability and validity in language test construction.

5.1 Language Testing in the Libyan EFL Context

Recent studies report ongoing discrepancies between the goals of Libyan English-language exams and their actual content. Onaiba (2024) carried out a document analysis of a high-stakes public EFL examination in Libyan schools and found that the examination was heavily based on decontextualised grammar items and ignored listening and speaking altogether, leading to the conclusion that present examination systems could impede curriculum implementation. Likewise, a 2022 study on the content validity of Tripoli public schools found a major gap between the content of the tests and the objectives of the curricula, raising questions about the fairness and effectiveness of the tests that are commonly used.

Elmahjoub (2023) offers a critical depiction of Libyan education as an exam-oriented culture, which he argues has impeded learner autonomy and constrained the scope of teaching. At the tertiary level, Elbouri (2022) examined the writing issues of Benghazi University English majors, and discovered that recurrent problems were frequent grammatical errors, poor organization and limited vocabulary.

Alsied (2020) found that Sebha University EFL students predominantly relied on memorization strategies that did little to support durable grammatical competence. Institutional reviews corroborate these findings. A 2020 report by the Democratic Arab Center identified resource shortages, as well as faculty's outdated pedagogies and lack of training, as structural barriers to effective EFL delivery. A related 2020 study revealed dissatisfaction among both educators and students with prevailing assessment practices, and called for more authentic, learning-oriented tools.

Two other Libyan studies provide further detail on this picture. Abdelaty (2023) investigated instructors' use of a required EFL textbook in Libya and concluded that a strict, test-driven approach to the prescribed textbook allowed little opportunity for communicative adaptation, perpetuating the same discrete, content-bound orientation identified in local testing practice. Therefore, Hadia (2023) investigated Libyan EFL teachers' views on the application of task-based approaches and revealed that although teachers were mostly in favour of task-based instruction, they indicated that the existing assessment systems were a disincentive to the regular use of this approach in the classroom. This is directly relevant to the washback dynamics investigated in this study.

5.2 Washback and Its Effects on Teaching and Learning

Washback, the effect of assessment on instruction and learning, is a topic of ongoing empirical interest. Rahman et al (2023) reviewed research on how stakeholders and contextual variables mediate washback. They concluded that although no single test can determine classroom practice on its own, teachers are the main agents through which washback—positive or negative—is produced. Specifically, in the Libyan context, Sharef (2024) used a mixed-methods design to show that assessment practices in secondary schools significantly redirect teaching toward examination results at the expense of communicative competence, while Ali (2024) found that testing shapes student motivation, instructional decisions, and lesson planning in ninth-grade classrooms. Taken together, these studies suggest a consistent pattern: where tests emphasise discrete, decontextualised knowledge, instruction tends to narrow in response, reducing opportunities for the extended, communicative language use that curricula nominally prioritise.

Two other studies are examples of washback moving in either direction depending on test design and teacher engagement. Athiworakun and Adunyarittigun (2022) investigated an exit exam at a Thai university and found that the test's intentional alignment with the CEFR and teachers' awareness of its communicative rationale resulted in mostly positive washback, with instructors managing exam preparation alongside more general skill development, compared to the negative washback observed at Sulugh Campus, where alignment and teacher buy-in were largely missing. In a review of four decades of washback research, Rathnayake (2024) also concludes that washback is not an intrinsic property of testing per se but rather a function of the way tests are designed, understood and used by teachers, thus supporting the argument that reforming test blueprints and developing teacher capacity are prerequisites for transforming washback from negative to positive.

5.3 Assessment Literacy and Teacher Preparedness

More and more research has shown a connection between the validity of classroom-based assessment and teachers' own assessment literacy. Chang et al. (2024) found that there were still gaps between the assessment competencies senior high school English teachers were expected to demonstrate and the training they actually received, a finding that is in line with broader work on language assessment literacy across EFL contexts (Allahyari et al., 2023). Research exploring alternatives to traditional discrete-point testing – such as portfolio assessment (Al-Hawamdeh et al., 2023), models of learner-oriented and self- or peer-assessment (Al-Rashidi et al., 2022), and reflective and formative techniques (Aldosari et al., 2023) – highlights improved communicative accuracy, fluency and

engagement in learners when teachers are willing to develop and interpret these approaches. In sum, this literature indicates that focused professional development on assessment literacy is a precondition for meaningful test reform, not an optional extra.

Regional evidence suggests similar gaps closer to the Libyan context. Elhussien and Khalil (2023) investigated the language assessment literacy of Saudi pre-service EFL teachers and found that they exhibited uniformly low levels of readiness for core testing tasks such as item writing and score interpretation, similar to the limited formal training reported by instructors in this study. In a similar vein, Tayyebi et al. (2022) explored classroom writing-assessment literacy among in-service EFL teachers in Iran, and reported that although teachers valued authentic writing tasks in principle, inconsistent application of rubrics and limited training in scoring productive skills compromised the validity of their classroom-based writing assessments – a pattern that resonates strongly with the scoring-consistency reservations expressed by instructors at Sulugh Campus.

5.4 Reliability and Validity in Language Test Construction

In classical measurement theory, test score is a function of true score and measurement error. Reliability estimates, most popularly Cronbach's alpha, are used to quantify the proportion of variance in test scores that can be attributed to true systematic differences among test-takers rather than random error (Izah et al., 2023). Interpretive conventions widely accepted suggest that alpha coefficients near .70 are adequate for group-level, low-stakes decisions, whereas much higher thresholds are recommended before a test is used for individual, high-stakes judgements. However, reliability is a necessary but not sufficient condition for a sound test. Validity, and in particular construct validity, is concerned with whether a test actually captures the underlying ability that it purports to measure (Messick, 1989). For tests of English proficiency, however, content that is composed predominantly of discrete grammar and vocabulary items, no matter how reliably scored, cannot be assumed to reflect validly a learner's communicative competence. The analysis offered in this study is organised around the difference between reliability and validity.

Two more papers add to the theoretical grounding. Hove (2022) revisited the unified, consensus-based definition of validity outlined in the AERA/APA/NCME Standards and argued that validation research should be grounded in the particular purposes for which test scores will be used, rather than considered as an abstract, one-off attribute of the test itself — a principle directly relevant to locally developed tests whose scores are used for both course progression and wider certification. In addition, Roebianto et al. (2023) provide a comprehensive procedural account of content validation, which emphasises systematic expert judgement and representative item coverage as the main safeguards against skewed, discrete-point content such as that reported in the results of this study.

5.5 Research Gap and Positioning of the Present Study

The above reviewed studies reveal important weaknesses in EFL assessment, but they treat these weaknesses in separate strands. Research in Libya has focused primarily on the content of the examination, teachers' or students' perceptions, classroom washback or the wider testing culture. International studies have provided evidence on assessment literacy, alternative assessment and validation procedures. The most directly relevant studies are synthesised in Table 1, which also shows the specific gap that is addressed by the present study.

Table 1. Summary of Previous Studies and the Research Gap Addressed by the Present Study

Study	Context / Focus	Main Finding Reported in the Literature Review	Gap Relative to the Present Study
Onaiba (2024)	Document analysis; high stakes EFL examination; Libyan school-level	The examination was largely based on decontextualised grammar items and excluded speaking leading to a mismatch between the aims of the examination and the aims of the curriculum .	The focus was on the exam content at the school level. It did not mix the reliability of the university test scores with the perceptions of stakeholders and classroom observation.
Tripoli content-validity study (2022)	Language Tests in Public Schools in Tripoli; Validity of Content	Found a major mismatch between the test content and the goals of the curriculum, raising issues of equity and effectiveness .	It focused on content validity and not reliability, construct validity, washback and stakeholder perceptions jointly.
Elmahjoub (2023)	Critical review of English language testing culture in Libya	Described an exam-centric culture in which excessive reliance on testing can narrow teaching and constrain learner autonomy.	Provided a broad critical account rather than an institution-specific empirical triangulation of scores, tests, interviews, focus groups, and observations.
Hadia (2023)	Libyan EFL teachers and task-based instruction	Teachers generally valued task-based approaches, but current assessment structures discouraged their consistent classroom use.	Examined teachers' perceptions of task-based teaching; it did not directly estimate test reliability or analyze the validity of a specific university testing system.
Sharef (2024) and Ali (2024)	Libyan secondary-school assessment and washback	Testing redirected teaching toward examination results and shaped motivation, instructional decisions, and lesson planning, often at the expense of communicative competence.	Focused on secondary-school contexts; the present study extends washback analysis to university-level locally developed proficiency tests and links it to psychometric evidence.
Chang et al. (2024); Elhussien & Khalil (2023)	Teacher language-assessment literacy in non-Libyan / regional EFL settings	Found gaps between expected assessment competencies and teachers' preparation, including low preparedness for item writing and score interpretation.	Demonstrated assessment-literacy problems but did not connect them to the reliability, validity, and washback of locally developed tests in a Libyan university.
Tayyebi et al. (2022)	In-service EFL teachers' writing-assessment literacy	Authentic writing was valued, but inconsistent rubric use and limited scoring training weakened assessment validity.	Focused on writing assessment; the present study examines a broader proficiency-testing system across multiple skills and data sources.
Athiworakun & Adunyarittigun (2022)	CEFR-aligned university exit examination in Thailand	A communicatively aligned test, supported by teacher understanding, produced largely positive washback.	Offers a useful contrast but comes from a different national context; the present study tests whether comparable alignment is present in a Libyan university context.
Roebianto et al. (2023)	Procedures for content validation	Emphasized systematic expert judgment and representative item coverage as safeguards for content validity.	Provides methodological guidance rather than empirical evidence on Libyan locally developed EFL tests.

Collectively, the reviewed literature therefore leaves a clear methodological and contextual gap. Within the studies reviewed for this paper, no single Libyan higher-education study combines internal-consistency estimates from actual locally developed test scores with direct analysis of test

content, instructor interviews, student focus groups, and classroom observations in order to evaluate reliability, construct/content validity, stakeholder perceptions, and washback within the same institutional case. The present study is designed to address that gap at Sulugh Campus.

5.6 Conceptual Framework

To synthesise the above-reviewed literature and to clarify the analytical logic of the study, a refined conceptual framework that guided the inquiry is presented in Figure 1. The phenomenon of interest is the locally developed English proficiency tests at Sulugh Campus and the analysis is organised around four interrelated dimensions, namely, reliability, validity, stakeholder perceptions, and washback.

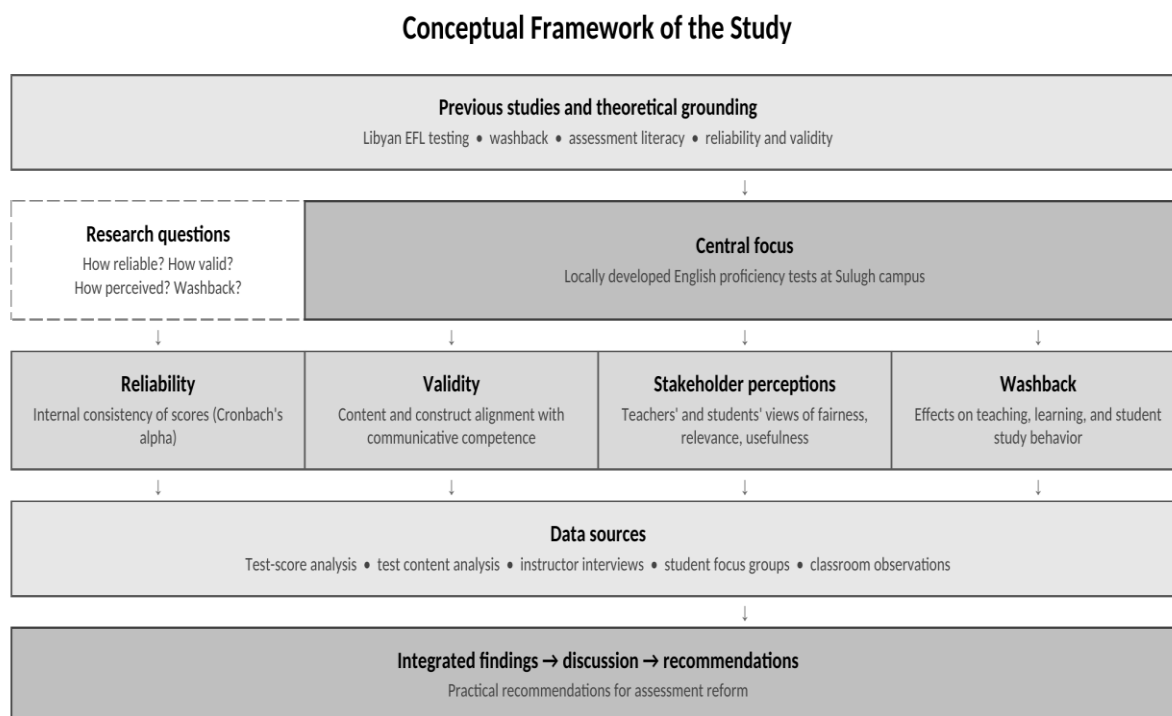


Figure 1. Refined conceptual framework of the study.

As shown in Figure 1, the framework draws on previous studies and theoretical grounding in language testing to structure the study around four core analytical dimensions. Reliability concerns the internal consistency of scores; validity addresses the extent to which the tests reflect communicative English proficiency rather than isolated language knowledge; stakeholder perceptions capture teachers' and students' views of fairness, relevance, and usefulness; and washback refers to the influence of testing on classroom practice and student learning behavior.

The framework also shows how these dimensions were examined through multiple data sources, namely test-score analysis, test content analysis, instructor interviews, student focus groups, and classroom observations. Taken together, these strands of evidence support an integrated interpretation of the findings and provide the basis for the discussion and practical recommendations proposed later in the study.

6. Research Methodology

6.1 Research Design

Guided by the conceptual framework presented in Figure 1, this study adopted a qualitative case study design supplemented by descriptive quantitative analysis, in order to examine the reliability and validity of locally developed English tests in the Department of English at Sulugh Campus, University

of Benghazi. The case study was situated at the Sulugh Campus in Suluq, where departmental tests serve as the principal mechanism for tracking student achievement and certifying proficiency.

6.2 Participants and Setting

Data were collected from two academic cohorts (2021/2022 and 2022/2023) enrolled in four core courses in the English Proficiency and Writing Skills sequences. Qualitative participants were five instructors involved in test development and delivery and twenty undergraduate students who participated in two focus groups.

6.3 Data Collection

Data collection was triangulated with four sources. First, the anonymised score sheets of the locally developed final examinations were analysed for internal consistency (Cronbach's alpha) and descriptive statistics. Second, the content, format and type of items and skill coverage of the examination were examined. Thirdly, semi-structured interviews were conducted with the five instructors to discuss test development, content selection and the challenges of assessing students in current conditions. Fourth, two focus groups with twenty undergraduates examined perceptions of fairness, relevance, difficulty, and clarity. These were supplemented by ten non-participant classroom observations, using a structured checklist to capture pedagogy, patterns of interaction and the degree of correspondence between classroom activities and test formats.

6.4 Data Analysis

Quantitative data was summarised using descriptive statistics and Cronbach's alpha coefficients. Qualitative data were collected through interviews, focus groups and observations and analysed thematically. The phases of thematic analysis described by Braun and Clarke (2006) were followed, with a combination of inductive data-driven coding and deductive codes derived a priori from the constructs of reliability, validity, washback and assessment literacy.

6.5 Ethical Considerations

Formal institutional approval was obtained before data collection. All instructor and student participants provided informed consent and anonymity was preserved throughout data handling, storage and reporting.

6.6 Trustworthiness and Limitations

As a single-site case study with a relatively small qualitative sample, this research does not claim statistical generalizability beyond the Sulugh Campus context. Transferability was instead pursued through thick description and methodological triangulation across quantitative, interview, focus-group, and observational data sources, consistent with recommended practice for context-bound case study research.

7. Results

7.1 Reliability of Local Tests

Internal consistency across the two cohorts was moderate, with Cronbach's alpha ranging from .65 to .75 across the four core courses examined (Table 2). These coefficients support interpretation at the group level but are not sufficiently robust to justify high-stakes decisions about individual students. Instructors reported that scoring of productive skills was considerably more variable — reflecting the difficulty of applying rubrics consistently — whereas they expressed greater confidence in the consistency of objective item formats such as multiple-choice and fill-in-the-blank questions.

Table 2. Internal Consistency (Cronbach's Alpha) by Cohort

Cohort	Courses Analyzed	Cronbach's Alpha (α)	Interpretation
2021/2022	English Proficiency & Writing Skills sequence (4 core courses)	.72	Moderate — acceptable for group-level interpretation
2022/2023	English Proficiency & Writing Skills sequence (4 core courses)	.68	Moderate — acceptable for group-level interpretation
Combined range	—	.65–.75	Moderate; insufficient for high-stakes individual decisions

7.2 Content and Construct Validity

Analysis of test content, corroborated by instructor and student accounts, indicated that discrete grammar and vocabulary items dominated the mark scheme, with reading assessed largely through short, literal comprehension tasks and writing assessed through short, accuracy-oriented tasks. Authentic speaking assessment was, by contrast, virtually absent (Table 3). Instructors attributed this imbalance to decontextualized item design, which they associated with construct-irrelevant variance — including test-wiseness and rote memorization — and to structural constraints such as large class sizes and limited training in assessment design.

Table 3. Approximate Distribution of Assessment Weight by Skill/Item Type

Skill / Item Type	Approximate Share of Marks	Format	Validity Concern Raised
Discrete grammar & vocabulary	≈60–70%	Multiple-choice, fill-in-the-blank	Construct-irrelevant variance (test-wiseness, memorization)
Reading comprehension	Minority share	Short, literal-level items	Limited coverage of inferential/critical reading
Writing	Minority share	Short, accuracy-oriented tasks	Extensive/contextualized writing under-represented
Speaking	Minimal to none	Largely absent from formal assessment	Communicative competence not assessed

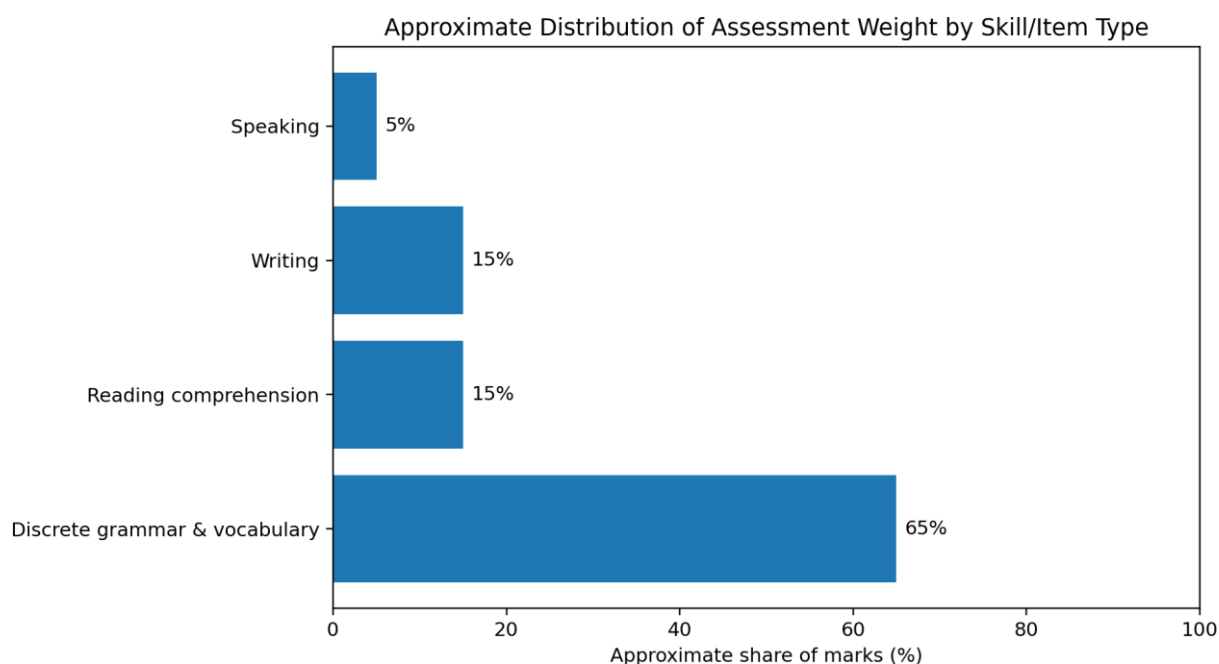


Figure 2. Approximate distribution of assessment weight by skill/item type.

Figure 2 is a visual approximation of mark distribution, on top of the descriptions reported in the results with discrete grammar and vocabulary items dominating and the marginal place of speaking, as a complement to Table 3.

7.3 Perceptions of Educators and Learners

Interview and focus group data indicated several common themes which are summarised in Table 4. Instructors reported limited formal training in testing and validation, though they generally found the exams necessary and feasible to administer, and questioned their ability to capture real communicative ability. For their own part, students preferred graded presentations and discussions rather than discrete items, partly because of additional test stress and partly because of a perceived mismatch between the content of the test and their actual communication needs.

Table 4. Summary of Qualitative Themes

Theme	Primary Source(s)	Key Finding
Limited assessment training	Instructor interviews	Instructors report minimal formal preparation in test design and validation
Preference for performance-based tasks	Student focus groups	Students favor presentations/discussions over discrete-item formats; report higher stress with current tests
Perceived construct mismatch	Instructors & students	Both groups view tests as misaligned with communicative goals and CEFR-based expectations
Structural constraints	Instructor interviews	Large class sizes and workload cited as barriers to more authentic assessment

7.4 Washback on Classroom Practice

Classroom observations indicated pervasive “teaching to the test,” with instruction dominated by past-format drills and grammar/vocabulary practice, particularly in the weeks preceding examinations (Table 5). Communicative activities received comparatively little classroom time, which observers linked to a narrowing of the curriculum and to negative washback: students' own study strategies shifted toward memorization rather than the development of fluency and strategic communicative competence.

Table 5. Classroom Observation Summary — Washback Indicators

Indicator	Observed Pattern	Washback Direction
Dominant activity type	Grammar/vocabulary drills and past-paper practice, especially pre-exam	Negative — narrows instructional focus
Communicative tasks (speaking/extended writing)	Infrequent; limited class time allocated	Negative — under-develops assessed-but-untested skills
Alignment of classroom activity with test format	High — instruction closely mirrors item types on exams	Negative — reinforces test-taking strategy over language development
Student study behavior (reported)	Shift toward memorization strategies	Negative — discourages transferable language learning

7.5 Comparison of the Present Findings with Previous Studies

To position the empirical findings in relation to the evidence reviewed in Section 5, Table 6 provides a descriptive comparison between the present study and the most directly relevant previous studies. The table is intended to show areas of convergence, extension, and contrast without replacing the fuller interpretation presented in the Discussion section.

Table 6. Comparison of Previous Studies with the Findings of the Present Study

Dimension	Present Study Finding	Relevant Previous Evidence	Relationship
Reliability	Cronbach's alpha ranged from .65 to .75, indicating moderate internal consistency that is more appropriate for group-level than high-stakes individual interpretation.	The reviewed Libyan studies largely focus on content, perceptions, or washback rather than reporting comparable internal-consistency estimates. Izah et al. (2023) describe alpha values around .70 as commonly adequate for group-level interpretation.	Extends the reviewed Libyan evidence by adding an explicit psychometric reliability estimate.
Content validity	Approximately 60–70% of marks were assigned to discrete grammar and vocabulary items, while speaking was minimal or absent.	Onaiba (2024) similarly reported decontextualized grammar emphasis and omission of speaking; the 2022 Tripoli study reported mismatch between test content and curricular aims.	Strong convergence with previous Libyan evidence.
Construct / communicative validity	The tests were perceived as poorly aligned with communicative competence and CEFR-oriented expectations; extended writing and spontaneous speaking were under-represented.	Elmahjoub (2023) described an exam-centric culture that narrows teaching; Hadia (2023) reported that assessment structures discouraged task-based communicative teaching.	Converges with prior concerns and demonstrates them directly within a university testing system.
Assessment literacy and scoring	Instructors reported limited formal preparation in test design and validation, with greater inconsistency when scoring productive skills.	Chang et al. (2024) and Elhussien and Khalil (2023) found gaps in assessment preparation; Tayyebi et al. (2022) linked inconsistent rubric use and limited scoring training to weaker validity.	Converges with regional/international evidence and applies it to the Sulugh Campus context.
Washback	Teaching was dominated by grammar/vocabulary drills and past-format practice; communicative tasks were infrequent and student study behavior shifted toward memorization.	Sharef (2024) found examination-driven teaching at the expense of communicative competence; Ali (2024) reported effects on motivation, instructional decisions, and lesson planning; Rahman et al. (2023) identifies teachers as key mediators of washback.	Strong convergence with previous washback research.

Dimension	Present Study Finding	Relevant Previous Evidence	Relationship
Positive contrast / reform potential	Current tests generated predominantly negative washback because communicative alignment and teacher preparation were limited.	Atthiworakun and Adunyarittigun (2022) reported largely positive washback when a university exit test was CEFR-aligned and teachers understood its communicative rationale.	Contrasts with a well-aligned model and indicates that washback may improve when test design and teacher capacity change.
Alternative assessment	Teachers and students expressed interest in more authentic and communicative assessment, including performance-oriented activities.	Al-Hawamdeh et al. (2023) and Al-Rashidi et al. (2022) reported benefits of portfolio, self-, peer-, and learner-oriented assessment for productive language outcomes and engagement.	Supports the practical direction suggested by stakeholder preferences in the present study.

As shown in Table 6, the present findings are broadly consistent with previous Libyan research concerning exam–curriculum mismatch, examination-driven teaching, and the under-representation of communicative skills. The principal added contribution of the present results is that these recurring qualitative concerns are observed alongside a direct internal-consistency estimate (.65–.75) and are triangulated across test-score data, test-content analysis, instructor interviews, student focus groups, and classroom observations. The comparison also shows that negative washback is not inevitable: evidence from a CEFR-aligned university examination indicates that better alignment and teacher understanding can support more positive classroom effects.

8. Discussion

Together these findings suggest that the locally constructed English tests at Sulugh Campus are moderately reliable but seriously lacking in validity. A Cronbach’s alpha of .65–.75 gives instructors reasonable confidence in the group-level interpretation of the objective sections. This reliability level does not, however, in itself imply that the tests measure the constructs they are designed to measure. Discrete grammar and vocabulary items are heavily favored, while integrated skills — spontaneous speaking, extended and contextualized writing — receive comparatively little attention. Such an imbalance creates a mismatch with goals of communicative competence and introduces construct-irrelevant sources of error, such as test-wiseness and rote memorisation.

Both instructor and student perspectives point to structural constraints – large class sizes, heavy workloads and insufficient training in assessment design – but also a clear appetite for more authentic and communicative assessment activities. The results are consistent with the larger research on washback. When teachers do not have the assessment literacy or resources to develop communicative tasks, negative washback tends to increase, resulting in curriculum narrowing, teaching to the test, and memorization-oriented study strategies that hinder critical thinking, fluency, and interactive communication (Chang et al., 2024; Rahman et al., 2023).

This pattern is consistent with other recent Libyan studies, including document-analysis evidence of exam–curriculum misalignment at the school level (Onaiba, 2024) and accounts of an exam-centric assessment culture more broadly (Elmahjoub, 2023) — suggesting the issues observed at Sulugh Campus are unlikely to be idiosyncratic to a single institution. At the same time, because this study is based on a single site and a relatively small qualitative sample, its findings should be read as indicative of a broader regional pattern rather than as directly generalisable in a statistical sense.

The reliability finding adds an aspect that is relatively less represented in the Libyan studies reviewed. The observed alpha range (.65–.75) is consistent with the broad measurement guidance outlined by Izah et al. (2023), where values around .70 may be acceptable for group-level interpretation but should not be taken as evidence that the tests are valid for all intended uses. This distinction is especially

important given Hoeve's (2022) argument that validation needs to be linked to the purposes for which scores are used. Thus, moderate internal consistency at Sulugh Campus could permit a degree of cautious interpretation of scores at the cohort level, but would need to be supplemented with stronger evidence, such as inter-rater reliability for productive skills and more detailed item-level analysis, to support consequential individual decisions with the same scores.

The content and construct-validity results are especially strongly consistent with previous Libyan evidence. Onaiba (2024) found that a high-stakes Libyan EFL examination emphasised decontextualised grammar, while ignoring speaking, and the 2022 Tripoli content-validity study reported a marked mismatch between assessment content and curricular aims. This pattern is replicated at university level in the present study: discrete grammar and vocabulary account for around 60-70% of marks, authentic speaking is virtually non-existent and extended writing is under-represented. This convergence means that the problem is not limited to one stage of education. It further supports the call for systematic content validation and representative item coverage by Roebianto et al. (2023) to avoid the under-representation of the construct.

The washback evidence is also in close agreement with previous research. Sharef (2024) found that the Libyan secondary-school assessment changed teaching to focus on examination results at the expense of communicative competence; Ali (2024) found that testing affects motivation, instructional decisions, and lesson planning. The same mechanism was evident in classroom observations at Sulugh Campus. Grammar/vocabulary drills and past-format practice increased prior to examinations, and communicative tasks were limited. Student study strategies shifted toward memorisation. This pattern is consistent with Rahman et al. (2023) who identify teachers and contextual constraints as important mediators of washback. Athiworakun and Adunyarittigun (2022) provide a contrasting case where CEFR alignment and teacher understanding were associated with largely positive washback. The contrast suggests that the negative effects identified here are better understood as products of test design, alignment, and implementation rather than as inevitable effects of testing per se.

The findings on assessment literacy offer a plausible institutional account for the weaknesses in validity and the inconsistency in scoring. Instructors in the present study reported limited formal preparation in test design and validation and more difficulty with consistently scoring productive skills. Similar gaps have been documented by Chang et al. (2024) among senior-high English teachers and by Elhussien and Khalil (2023) among pre-service EFL teachers. Moreover, inconsistent use of rubrics and limited scoring training can threaten the validity of writing assessment (Tayyebi et al., 2022). These parallels strengthen the argument for systematic assessment-literacy development that emphasises blueprint construction, task design, rubric development, standardisation, moderation, and score interpretation that are not one-off workshops.

Finally, stakeholder interest in presentations, discussions and other authentic tasks is consistent with evidence on alternative assessment. Al-Hawamdeh et al. (2023) and Al-Rashidi et al. (2022) reported positive effects of portfolio and learner-oriented assessment approaches on productive language outcomes, engagement, and other learner variables. These studies are not a direct intervention test for Sulugh Campus, and their findings should not be interpreted as evidence that the same effects will be observed locally. But they provide an evidence-based rationale for piloting a more balanced system in which authentic speaking and extended writing tasks, formative assessment, and consistent analytic rubrics complement rather than just replace efficiently scored objective items.

Overall, the current study confirms and extends the previous literature. It confirms recurrent Libyan concerns relating to content imbalance, communicative under-representation, exam-driven teaching and limited assessment preparation; it extends them by bringing together these concerns with internal-consistency evidence and four-source qualitative triangulation within a single university department.

This integrated design is the main contribution in relation to the research gap identified in Section 5.5 and offers a stronger basis for linking test quality to classroom consequences than any single source of evidence would on its own.

9. Implications and Recommendations

Based on these findings, four priority actions are recommended for the Department of English at Sulugh Campus and comparable institutions:

- Revise test blueprints to incorporate authentic communicative tasks, including structured speaking assessment, aligned with CEFR descriptors of communicative competence.
- Invest in systematic, ongoing professional development in assessment literacy for instructors, addressing both test design and the consistent application of scoring rubrics for productive skills.
- Integrate formative assessment approaches — including portfolio-based and self-/peer-assessment models — to complement summative testing and to encourage continuous, communicatively oriented learning (Al-Hawamdeh et al., 2023; Al-Rashidi et al., 2022).
- Establish continuous validation cycles, in which test content and format are periodically reviewed against curricular aims and stakeholder feedback, to prevent the kind of exam–curriculum drift documented in comparable Libyan contexts (Onaiba, 2024).

10. Limitations and Directions for Future Research

This study is limited by its single-site design and comparatively small qualitative sample, which constrain statistical generalizability beyond Sulugh Campus. The reliability analysis was restricted to Cronbach's alpha; future research could usefully incorporate additional reliability estimates (e.g., inter-rater reliability for productive-skill scoring) and more fine-grained item-level analysis. Future studies would also benefit from multi-site designs across Libyan universities, longitudinal tracking of washback following test-reform interventions, and direct empirical measurement of the relationship between instructors' assessment literacy and the validity of the tests they design.

11. Conclusion

The findings of this case study confirm that locally developed English proficiency tests in Sulugh Campus have moderate reliability (Cronbach's alpha .65-.75) and significant validity deficits. The tests favour discrete knowledge of grammar and vocabulary and fail to test integrated skills of extended writing, spontaneous speaking and interactive communication. Classroom observations and stakeholder accounts suggest widespread negative washback, with instruction focused on test preparation and student learning strategies orientated toward rote memorisation rather than fluency and interaction. Among top priority areas for improvement are increasing authentic speaking and writing assessment, revising test blueprints to more communicative and CEFR-aligned models, and investing in targeted professional development to strengthen instructors' assessment literacy, rubric use, and validation practices. A more balanced system of assessment, combining formative and summative approaches, has the potential to improve learning outcomes, reduce pressure around testing and increase fairness. Continued stakeholder feedback will be necessary to maintain the relevance, fairness, and usefulness of English-language assessment at Sulugh Campus and similar Libyan institutional contexts.

References

Abdelaty, S. (2023). Teachers' use of a compulsory EFL textbook in Libya: A mixed methods study. *Biomedical Journal of Scientific & Technical Research*, 50(2), 41586–41591.

- Aldosari, M., Heydarnejad, T., Hashemifardnia, A., & Abdalgane, M. (2023). The interplay among self-assessment, using reflection for assessment, classroom enjoyment, and immunity: Into prospects of effective language learning. *Language Testing in Asia*, 13, Article 1. <https://doi.org/10.1186/s40468-022-00213-1>
- Algwil, K. A. (n.d.). *Libyan EFL teachers' beliefs about learner autonomy at a Libyan university* [Unpublished manuscript].
- Ali, H. (2024). Washback in English education: Classroom impacts of testing in Libyan ninth grade. *Journal of Educational Studies*, 12(1), 45–62.
- Al-Hawamdeh, B. O. S., Hussien, N., & Abdelrasheed, N. S. G. (2023). Portfolio vs. summative assessment: Impacts on EFL learners' writing complexity, accuracy, and fluency (CAF); self-efficacy; learning anxiety; and autonomy. *Language Testing in Asia*, 13, Article 12. <https://doi.org/10.1186/s40468-023-00225-5>
- Allahyari, R., Abbasabady, M. M., Akhter, S., & Alibakhshi, G. (2023). EFL teachers' cognition of social and psychological consequences of high-stakes national language tests: The role of teacher training workshops. *Language Testing in Asia*, 13(1). <https://doi.org/10.1186/s40468-023-00262-0>
- Al-Rashidi, A. H., Asif, M., Vanani, M. G., & Aberash, A. (2022). Learner-oriented assessment (LOA) practice: A comparative study of self-, peer, and teacher assessment on EFL learners' writing and speaking CAF and attitude. *Language Testing in Asia*, 12.
- Alsied, S. (2020). Grammar learning strategies among Libyan EFL undergraduates at Sebha University [Working paper]. Semantic Scholar.
- Athiworakun, C., & Adunyarittigun, D. (2022). Investigating washback effects on teaching: A case study of an exit examination at the higher education level. *LEARN Journal: Language Education and Acquisition Research Network*, 15(2), 776–800.
- Baba, A. O. (2025). The impact of using L1 pedagogical practices in English as foreign language (EFL) classrooms: A case study of Emhimmed Ben Ebrahim Secondary School teachers. *مجلة الأكاديمية الليبية بنى وليد (JLABW)*, 392–400.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Chang, D. Y.-S., Lin, M. H., & Lee, J.-Y. (2024). Exploring language assessment literacy and needs of English teachers at senior high school level. *Asia Pacific Journal of Education*, 44(4), 854–872. <https://doi.org/10.1080/02188791.2024.2313486>
- Democratic Arab Center. (2020). *Teaching and learning EFL in Libyan higher education: Challenges and implications*. Democratic Arab Center Publications.
- Elbouri, A. (2022). Writing problems among English majors at the University of Benghazi: A case study. *Research Journal of Language, Literature and Humanities*, 9(2), 10–18.
- Elhussien, F. A., & Khalil, S. M. (2023). Examining language assessment literacy for Saudi pre-service EFL teachers. *World Journal of English Language*, 13(6), 431–438. <https://doi.org/10.5430/wjel.v13n6p431>
- Elmahjoub, A. A. (2023). A testing-laden culture: A critical review of English language testing in the Libyan education system. *Libyan Journal of Educational Sciences*, 8, 233–246.
- Hadia, G. (2023). EFL Libyan teachers' perceptions of implementing task-based approach into their teaching. *Bani Waleed University Journal of Humanities and Applied Sciences*, 8(4), 100–119.
- Hoeve, E. (2022). A validity framework for accountability: Educational measurement and language testing. *Language Testing in Asia*, 12, Article 3. <https://doi.org/10.1186/s40468-022-00153-w>

- Investigating language tests' content validity in public schools in Tripoli. (2022). *ResearchGate* [Preprint].
- Izah, S. C., Sylva, L., & Hait, M. (2023). Cronbach's alpha: A cornerstone in ensuring reliability and validity in environmental health assessment. *ES Energy & Environment*, 23, Article 1057.
- Latif, M. W., & Wasim, A. (2024). Investigating EFL instructors' approaches to classroom-based assessment culture: An explanatory sequential mixed-method approach. *Language Testing in Asia*, 14, Article 57. <https://doi.org/10.1186/s40468-024-00332-x>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.
- Noroozi, S., & Karami, H. (2024). A Rasch-based validation of the University of Tehran English Proficiency Test (UTEPT). *Language Testing in Asia*, 14, Article 18. <https://doi.org/10.1186/s40468-024-00290-4>
- Onaiba, A. (2024). Do exam aims and content reflect those of the curriculum? An evaluative study. *Asian Journal of Assessment in Teaching and Learning*.
- 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). <https://doi.org/10.55593/ej.28110a8>
- Rahman, K. A., Rukanuddin, M., Chowdhury, M. S. Y., Ahmed, S., & Seraj, P. M. I. (2023). Recognizing stakeholders and factors mediating washback in language testing. *Education Research International*, 2023, Article 5548723. <https://doi.org/10.1155/2023/5548723>
- Rathnayake, W. M. P. Y. B. (2024). From tests to teaching: Exploring the washback effect in language pedagogy. *The Journal of Desk Research Review and Analysis*, 2(2), 145–158. <https://doi.org/10.4038/jdr.v2i2.47>
- Roebianto, A., Savitri, S. I., Aulia, I., Suciñana, A., & Mubarakah, L. (2023). Content validity: Definition and procedure of content validation. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, 30(1), 5–18.
- Sharef, M. (2024). Exploring the pedagogical impact of testing in Libya: A mixed-methods study of secondary EFL assessment practices. *Journal of Research and Practice in Education*, 5(9), 1–18.
- Sureeyatanapas, P., Sureeyatanapas, P., Panitanarak, U., Kraisiwattana, J., Sarootyanapat, P., & O'Connell, D. (2024). The analysis of marking reliability through the approach of gauge repeatability and reproducibility (GR&R) study: A case of English-speaking test. *Language Testing in Asia*, 14, Article 1. <https://doi.org/10.1186/s40468-023-00271-z>
- Tayyebi, M., Abbasabady, M. M., & Abbassian, G. R. (2022). Examining classroom writing assessment literacy: A focus on in-service EFL teachers in Iran. *Language Testing in Asia*, 12(1). <https://doi.org/10.1186/s40468-022-00161-w>
- Zhang, H. (2024). The washback of the National Matriculation English Test on senior high school English learning outcomes: Do test takers from different provinces think alike? *Language Testing in Asia*, 14, Article 14. <https://doi.org/10.1186/s40468-024-00286-0>