

Language Assessment Literacy and Classroom-Based Language Testing Practices among EFL Teachers in the Riyadh Region, Saudi Arabia

Rami Mubarak

Department of Foreign Languages, Taif University, Taif, Saudi Arabia

Approved: 06 August 2026

Posted: 08 August 2026

Copyright 2026 Author(s)

Under Creative Commons CC-BY 4.0

OPEN ACCESS

Cite As:

Mubarak, R. (2026). *Language Assessment Literacy and Classroom-Based Language Testing Practices among EFL Teachers in the Riyadh Region, Saudi Arabia*. ESI Preprints. <https://doi.org/10.19044/esipreprint.8.2026.p69>

Abstract

In applied linguistics, language assessment literacy has emerged as a fundamental construct denoting teachers' knowledge, skills, and principles concerning language assessment. In EFL contexts, where classroom-based language testing is mostly designed and administered by teachers, it is crucial to understand how assessment literacy relates to testing practices. This cross-sectional correlational study examined Language Assessment Literacy (LAL), classroom-based language testing practices, and the relationship between the two constructs among 130 EFL teachers from six secondary schools and four universities in the Riyadh Region of Saudi Arabia. Of 168 eligible teachers invited between October 6 and November 28, 2025, 142 submitted questionnaires (84.5%), and 130 records met the analysis criteria. Participants completed a 25-item Language Assessment Literacy Questionnaire and a 20-item Classroom-Based Language Testing Practices Scale. Both researcher-developed instruments were informed by established frameworks, reviewed by five experts, and pilot-tested with 18 Saudi EFL teachers. Overall LAL was moderate ($M = 3.62$, $SD = 0.47$, $\alpha = .88$), and reported testing practices were also moderate ($M = 3.57$, $SD = 0.44$, $\alpha = .86$). Overall LAL correlated positively with reported practices, $r(128) = .63$, $p < .001$, 95% CI [.51, .72]. The regression model explained 39.7% of practice-score variance, $R\text{-squared} = .397$, adjusted $R\text{-squared} = .392$, $F(1, 128) = 84.24$, $p < .001$. Because both constructs were measured

through self-report, the findings establish association rather than causation or verified classroom performance.

Keywords: Language assessment literacy; classroom-based language testing; applied linguistics; EFL teachers; Saudi Arabia; language measurement

1. Introduction

1.1 Problem introduction

Assessment constitutes a central component of language measurement practices in English as a Foreign Language (EFL) contexts since it fulfils the functions of measuring language proficiency, interpreting language performance, and informed use of assessment outcomes in making decisions related to testing. Classroom-based language assessment goes beyond a terminal evaluation role and serves as a continuous language measurement process, score interpretation and decision-making in local testing settings. (Sujianti et al., 2025). Classroom-based language testing, specifically, has become one of the key elements of assessing the progress and the success of the students, since it is usually the most immediate and frequent type of assessment that learners in EFL environments experience (Tewachew, 2025).

In most of the EFL cases, instructors perform the role of teachers and assessors. Their duty will be to design tests, choose assessment activities, grade the performance of students, and interpret the results to make pedagogical choices. In comparison to large-scale standard evaluation, classroom-based assessment is mostly teacher-oriented and situation-specific meaning it puts significant pressure on the professional judgement of teachers. Consequently, the quality of language testing in the classroom is directly connected with the knowledge of the teacher in the principles of assessment and his/her skill to implement the principles in a proper way (Phakiti and Steinhoff, 2023).

Nevertheless, studies have always questioned the quality of the classroom-based testing of language in EFL. Poor assessment practices, including poorly designed test items, excessive reliance on discrete-point testing, and a lack of correspondence between instructional goals and assessment activities, have been widely reported. These practices can lead to misrepresentation of the targeted language constructs and invalidity of the assessment outcomes (Mahar and Khoso, 2025). In addition, the lack of proper assessment practices may negatively affect perceptions of assessment fairness, score credibility, and the interpretability of assessment outcomes.

Language Assessment Literacy (LAL) is an idea that has been adopted in the field of applied linguistics to help overcome these issues. LAL

entails knowledge, skills, and principles of teachers in regards to language assessment which includes knowledge of test validity and reliability, scoring procedures, feedback, and ethical implications. Teachers who possess adequate amounts of LAL will be in a better position to design and develop assessments, which aptly capture the language competence of learners and foster purposeful learning (Tsagari & Rousoulioti, 2025). Ineffective LAL, on the other hand can lead to inconsistent, subjective and misaligned assessment practices that do not meet curricular goals.

In most EFL situations, a teacher has limited formal exposure to principles of language assessment within their professional trajectories. This has led to a situation where decisions about assessment are based on personal experience, institutional patterns or the textbooks that are available instead of sound assessment principles. These practical issues raise the problem of the necessity of exploring the Language Assessment Literacy of EFL teachers and the impact it has on classroom-based language testing practices.

The present study is bounded to the Riyadh Region of Saudi Arabia, where English is taught as a foreign language in centrally regulated secondary schools and institutionally diverse university programmes. Secondary teachers commonly work within Ministry-aligned curricula and school examination schedules, whereas university instructors may follow programme-level test specifications, common examinations, or course-coordinator requirements. These differences make educational level, institutional ownership, moderation procedures, and access to assessment training relevant to the interpretation of teachers' reported practices.

1.2 Explore the Importance of the Problem

The Language Assessment Literacy investigation is of paramount significance in the practice of linguistics since the notion of assessment functions as a central mechanism through which language ability is measured and assessment-related decisions are made. Assessment of language determines what will be taught, how the students will learn, and the definition of achievement in a learning context. Assessment practices that are based on good assessment literacy can bring about positive washback by aligning assessment tasks with targeted language constructs and intended test use (Green, 2007). On the contrary, poorly constructed assessments can produce negative washback, thus causing greater emphasis on test-taking strategies at the expense of construct representation and score validity.

LAL is of concern to the major assessment qualities like validity, reliability, fairness and transparency. Validity assures that an assessment measures what it is supposed to measure whereas reliability is the consistency of assessment outcomes (Gupta, 2023). Ethical assessment practices and fairness are necessary to help make sure that learners are

assessed fairly and without prejudice. The higher the LAL of the teachers, the more they will tend to consider such principles in drafting and conducting classroom-based testing, the more they will be able to improve the quality of the assessment outcomes and make them credible.

Although the assessment theory within applied linguistics has been expanding, there has been an outstanding gap separating the theoretical concepts from the practical implementation in the classroom. Although assessment frameworks and guidelines are well articulated in the literature they are not commonly applied practically in the conventional EFL classrooms. Teachers might have theoretical knowledge of assessment concepts without always converting this into assessment decision-making and testing behaviour. This loophole is quite pronounced in situations where assessment training is minimal or where high-stakes examinations prevail in the educational system.

The problem is especially acute in EFL settings where the possibility of systematic involvement in language assessment principles is restricted. In these environments, teachers can mostly depend on traditional testing measures, including multiple-choice items and grammar-based tests, that might not sufficiently represent targeted communicative language constructs. Language Assessment Literacy research into teacher practices can thus yield data on the factors that are underlying the poor classroom-based language testing practices.

The current study helps in understanding variation in the quality, consistency, and interpretability of classroom-based language assessment practices because it focuses on the connection between Language Assessment Literacy and classroom-based language testing practices. It offers empirical data that informs the knowledge about assessment-related competence and contributes to understanding variation in language assessment competence across EFL contexts.

1.3 *Describe Relevant Scholarship*

Language assessment literacy is a topic that has received a lot of debate in the area of applied linguistics within the last twenty years, especially in regard to the growing need for accountability and assurance of quality in language education. Initial conceptualisations of assessment literacy centred on a broad educational assessment, where the skill of teachers to interpret test outcomes and apply them correctly was regarded (Gareis and Grant, 2015). In more recent work, this idea has been applied to the contexts of specific languages, where it is acknowledged that there are particular issues connected with measuring language proficiency (Tsayari and Rousoulioti, 2025).

Major theorists have theorised LAL as a multidimensional concept which comprises evaluation knowledge, applied skills and underlying principles or values. Assessment knowledge incorporates the knowledge of concepts like validity, reliability, test formats, and scoring methods (Gupta, 2023). Assessment skills are the skills that allow designing assessment tasks, analysing results and providing effective feedback. The principles of assessment include ethical principles, fairness and awareness of the social impact of assessment. Collectively, the dimensions constitute a full picture in defining teacher competence in language assessment.

Empirical research on the Language Assessment Literacy of teachers has given inconclusive results. There are studies saying that teachers have a shallow understanding of assessment formats but do not have a deeper knowledge of what is meant by concepts like validity and reliability (Shafii and Berger, 2025). Other studies have detected inconsistencies between teachers' self-reported assessment knowledge and their relationship to their assessment decision-making and testing practices (Koloi-Keaikitse & Traynor, 2023). It has also been illustrated in research that contextual factors, including institutional policies and pressures of examination, influence assessment behaviours of teachers despite their assessment literacy levels (Prastikawati et al., 2025).

Saudi evidence reinforces the importance of this bounded context. Umer et al. (2018) reported gaps between assessment theory and practice among Saudi university EFL teachers, while Almosa and Alzahrani (2022) showed that institutional assessment arrangements influence the approaches available to higher-education teachers. More recently, Latif and Wasim (2024) found that traditional assessment remained prominent and that teachers' stated approaches did not always correspond with documented classroom practice. These studies justify examining a mixed group of secondary and tertiary teachers while avoiding the assumption that all EFL settings operate in the same way.

Research on classroom-based assessment practices has also determined chronic issues, such as the lack of alignment between assessment tasks and instructional goals, the lack of use of formative assessment, and the inadequate provision of feedback to learners (Bravo & Paceno, 2025). Such results indicate that the current assessment practices are usually imposed by practical limitations, but not by the pedagogical factors, which can make them less effective in facilitating language learning.

Although LAL has increasingly attracted attention, the literature gap on the empirical relation between the Language Assessment Literacy of teachers and the classroom-based language testing practices is very significant. The literature has considered LAL as a separate construct without necessarily relating it to visible assessment behaviours (Gal &

Ryder, 2025). Therefore, the empirical evidence available to pinpoint the way the assessment literacy of teachers is reflected in the real testing practices in EFL classrooms is scarce. This gap is critical to curb the area of research in applied linguistics as well as to devise evidence-based measures to enhance classroom assessment.

1.4 State Research Questions and Their Correspondence to Research Design

The main aim of the current research is to study the Language Assessment Literacy of EFL teachers in the Riyadh Region and to evaluate its relationship with classroom-based language testing practices. Through an empirical method, the research aims to give a systematic explanation of the evaluation competence of teachers and its application in Saudi secondary and tertiary EFL classrooms.

To achieve this purpose, the study addresses the following research questions:

1. What is the level of Language Assessment Literacy among participating EFL teachers in the Riyadh Region?
2. What dimensions of Language Assessment Literacy are most and least developed among participating EFL teachers?
3. What is the reported level of classroom-based language testing practices among participating EFL teachers?
4. Is there a significant relationship between teachers' Language Assessment Literacy and their classroom-based language testing practices?

Quantitative research design was used to answer these questions because it enables the systematic measurement of assessment literacy and testing practices by the teachers on a relatively large scale. This is possible using survey tools and statistical tests that help in detecting patterns, relationships, and changes in the data (Sheikh et al., 2024). This is a design which is especially suitable to investigate the strength and direction of the relationship between the Language Assessment Literacy and classroom-based language testing and therefore be able to offer empirical data to guide the applied linguistics research and practice.

2. Method

2.1 Research Design

The research design used in this study was a quantitative correlational research design, in which the authors attempted to establish a relationship between the Language Assessment Literacy (LAL) of EFL teachers and the language testing practices in his/her classroom. A correlational design would be suitable where the objective of the study is to test the strength and

direction of relationships among variables without interfering with them. Considering that the current research is aimed at understanding naturally occurring variations of the assessment literacy of teachers and the correlation of the variations with the assessment practices of the teachers, an experimental or quasi-experimental design was neither feasible nor suitable.

A survey-based data collection method was used as the main data collection method. Surveys have a long history of application in applied linguistics studies in terms of both obtaining data on fairly large sample sizes and in terms of quantifying research constructs, including knowledge, beliefs, and practices. When it comes to language assessment research, questionnaires can be used for systematic measurement of the assessment behaviours and knowledge of teachers, which can be statistically analysed. In addition, a survey design will provide uniformity in the data collection of all the participants and this will enable comparison of various demographic groups.

The study design had a direct relationship with the research questions. The first three study questions, the goal of which is to calculate the general level of LAL and identify its most and least developed dimensions, demanded descriptive statistics. The fourth research question that investigated the links between LAL and classroom-based language testing practices required correlational and inferential statistical procedures. In this regard, the quantitative correlational design offered a coherent scheme of addressing all the research questions using one methodological approach.

Language Assessment Literacy was the independent variable in this study and it was conceptualised as a multidimensional construct that includes knowledge, skills, and principles of teachers related to language assessment. The dependent variable was the classroom-based language testing practices which represented the way in which teachers design, administer, score, and use language assessments in their classrooms. The connection of these two variables was investigated through the application of proper statistical methods to establish the hypothesis of whether a greater level of LAL is linked to more sound and principled practices of assessment.

The study was cross-sectional. Consequently, treating LAL as the predictor and reported testing practices as the outcome in the regression model indicates statistical prediction within this dataset; it does not establish temporal ordering or causation.

2.2 *Research Context and Participating Institutions*

The study was conducted in the Riyadh Region of Saudi Arabia. Ten institutions participated: six secondary schools and four universities. To protect institutional identity, the sites are reported by codes rather than names. S1-S4 were public secondary schools, S5-S6 were private secondary

schools, U1-U2 were public universities, and U3-U4 were private universities. Teachers worked in general English, foundation-year English, English-major, or English-for-specific-purposes courses.

Each participating institution nominated a coordinator who circulated an identical invitation but did not know which teachers completed the survey. Table 1 reports the contribution of each anonymised site to the final sample.

Table 1: Participating Institutions and Final Sample Contribution

Site code	Educational level	Ownership	n
S1	Secondary school	Public	14
S2	Secondary school	Public	13
S3	Secondary school	Public	12
S4	Secondary school	Public	10
S5	Secondary school	Private	12
S6	Secondary school	Private	11
U1	University	Public	20
U2	University	Public	17
U3	University	Private	11
U4	University	Private	10
Total	—	—	130

Note. Institutional codes protect site confidentiality. Counts describe the 130 records retained for analysis.

2.3 Participants

The EFL teachers who taught the English language as a foreign language in formal learning institutions were the target population of this study. The participants were recruited across different teaching settings such as secondary and tertiary education settings to provide variability in teaching experience and assessment duties. The choice of these contexts is due to the fact that teachers working in these settings are usually tasked with the design and implementation of classroom-based language tests and assessments.

The participant group comprised teachers of varied demographic backgrounds including gender, years of teaching experience, and academic qualifications. The overall range of expertise was between early-career teachers who had less than five years of teacher experience and more experienced teachers with more than ten years of classroom practice. Academic credentials ranged between bachelor's degrees in English or applied linguistics and postgraduate degrees, such as master's degrees in TESOL, applied linguistics or related fields. This variety made possible a more holistic representation of EFL teacher assessment literacy at the different levels of professionalism.

In order to achieve the relevance and validity of the data, multiple inclusion criteria were used. To begin with, the participants had to be teaching English as a foreign language at the moment at which the data was collected. Second, they had to be directly involved in the evaluation of the

language performance of students in classroom-based tests, quiz tests, or any other type of assessment. Third, the participants must have a minimum of one year of teaching experience to guarantee that they were exposed to classroom assessment practices. The teachers not engaged in assessment-related activity or those who only taught English in administrative or supervisory activities were not included in the study.

The final sample consisted of 130 teachers. Women comprised 57.7% of the sample, 55.4% taught at secondary level, 66.2% worked in public institutions, and 48.5% held a postgraduate qualification. Experience was distributed across all three categories. All participants worked in the Riyadh-region Saudi EFL context described above. Table 2 provides the complete demographic profile.

Table 2: Demographic Profile of Participants (N = 130)

Variable	Category	n	%
Gender	Women	75	57.7
	Men	55	42.3
Teaching level	Secondary	72	55.4
	Tertiary/university	58	44.6
Teaching experience	1-5 years	38	29.2
	6-10 years	47	36.2
	More than 10 years	45	34.6
Highest qualification	Bachelor's degree	67	51.5
	Master's degree	50	38.5
	Doctorate	13	10.0
Institution type	Public	86	66.2
	Private	44	33.8
Country/context	Saudi Arabia (Riyadh Region)	130	100.0

Note. Percentages are calculated within each demographic variable and may differ slightly from 100 because of rounding.

2.4 Sampling and Recruitment Procedures

In this study, the participants were recruited using a non-probability sampling approach. In particular, convenience sampling was applied because it is practical and accessible in the context of the research. Such a sampling approach is typical of studies in applied linguistics, especially when the researcher is limited in their access to participants due to institutional or other logistical reasons. Teachers who fulfilled the inclusion criteria and were interested in taking part were invited to fill in the questionnaire.

The participants were contacted via professional networks, institutional contacts and direct contact with the schools and universities. To make informed consent, information on the purpose and procedure of the study was given beforehand. Even though convenience sampling impairs the generalisability of the results, it is appropriate in the exploratory and

correlational research designed to reveal patterns and relationships without making any claims about the whole population.

Between October 6 and November 28, 2025, site coordinators sent a standardised electronic invitation to 168 eligible teachers. Two reminders were issued, 10 and 20 days after each site's initial invitation. A total of 142 questionnaires were submitted, producing a gross response rate of 84.5%. Twelve records were excluded: three repeated submissions identified through single-use anonymous codes, five respondents who did not meet the current-teaching or assessment-responsibility criteria, and four questionnaires with more than 20% of instrument items missing. The final 130 cases represented 77.4% of invited teachers and 91.5% of submitted questionnaires.

2.4.1 Sample Size, Power, and Precision

The sample size as a whole comprised 130 EFL teachers, which is within the suggested limit for correlational studies in social science research. This sample was deemed sufficient since it provided enough statistical power to determine moderate relationships between variables. Past methodological research indicates that a sample size of 100 or more is sufficient in carrying out a correlational study with questionnaire data especially when reliability and validity studies are to be undertaken.

An a priori calculation for a two-tailed Pearson correlation specified a moderate target association of $r = .30$, $\alpha = .05$, and $\text{power} = .80$. The minimum requirement was 84 cases; the target was increased to at least 120 to support dimension-level descriptive estimates and anticipated exclusions (Faul et al., 2009).

The necessary sample size was chosen since it allows for the meaningful descriptive statistics, internal consistency analysis, and inferential testing, i.e., Pearson correlation and simple regression analysis. Although the sample might not have incorporated all EFL teachers in all settings, it offers an adequate dataset in the study with regard to analysing the correlation between Language Assessment Literacy and classroom-based language testing practices within the context of this study.

2.5 Research Instruments

The quantitative data needed in this study were collected using two self-report questionnaires.

2.5.1 *Language Assessment Literacy Questionnaire*

The researcher-developed Language Assessment Literacy Questionnaire (LALQ) was informed by established language-assessment-literacy frameworks rather than copied verbatim from a previously validated instrument. It contained 25 items: five each for assessment validity, reliability, test design and scoring, feedback-related score use, and ethics and

fairness. Items assessed teachers' perceived ability to make language-assessment judgements in secondary and tertiary EFL settings.

Responses ranged from 1 (strongly disagree) to 5 (strongly agree). Two negatively worded items, R5 and F5, were reverse-scored. Each dimension score was calculated as the mean of its five items, and the overall LALQ score was the mean of all 25 items. Higher values indicated stronger perceived LAL.

2.5.2 Classroom-Based Language Testing Practices Scale

The researcher-developed Classroom-Based Language Testing Practices Scale (CBLTPS) contained 20 items: five each for construct alignment, test design practices, scoring consistency, and score interpretation and reporting. The items measured how frequently teachers reported using practices related to assessment planning, task design, scoring, feedback, interpretation, and reporting.

Teachers responded from 1 (never) to 5 (always). Item SC5 was reverse-scored. Dimension and overall scores were calculated as item means, with higher values indicating more frequent principled practice. For both instruments, means of 1.00-2.33 were interpreted as low, 2.34-3.67 as moderate, and 3.68-5.00 as high. These descriptive bands were specified before analysis and do not represent external performance standards.

Table 3: Instrument Structure, Sample Items, and Main-Sample Reliability

Scale/dimension	Items	Illustrative item	alpha
LALQ: Validity	5	I can judge whether an assessment represents the intended language ability.	.84
LALQ: Reliability	5	I can identify likely sources of inconsistency across tasks or raters.	.82
LALQ: Test design and scoring	5	I can construct tasks and scoring criteria that match learning outcomes.	.81
LALQ: Feedback-related score use	5	I can use result patterns to plan feedback and reteaching.	.79
LALQ: Ethics and fairness	5	I can identify features that may unfairly disadvantage learners.	.83
LALQ overall	25	Mean of all five dimensions	.88
CBLTPS: Construct alignment	5	Before writing a test, I map each task to a learning outcome.	.82
CBLTPS: Test design practices	5	I review items for ambiguity and unintended clues.	.80
CBLTPS: Scoring consistency	5	I use written criteria or rubrics for productive language.	.81
CBLTPS: Interpretation and reporting	5	I explain scores in relation to assessed learning outcomes.	.78
CBLTPS overall	20	Mean of all four dimensions	.86

Note. LALQ = Language Assessment Literacy Questionnaire; CBLTPS = Classroom-Based Language Testing Practices Scale. Complete items appear in Appendix A.

2.6 *Instrument Development, Reliability, and Validity*

The instruments were researcher-developed and framework-informed. Construct domains were mapped to work on LAL content and training needs (Fulcher, 2012; Vogt & Tsagari, 2014), situated teacher assessment literacy (Xu & Brown, 2016), language-specific assessment competence (Levi & Inbar-Lourie, 2020), and classroom assessment practice (Phakiti & Steinhoff, 2023). Five reviewers—three language-assessment specialists, one EFL teacher educator, and one educational-measurement specialist—rated relevance and clarity on a four-point scale. Item-level content-validity indices ranged from .80 to 1.00; average scale-level indices were .94 for the LALQ and .93 for the CBLTPS. Seven items were simplified, and two overlapping items were replaced before piloting.

The revised instruments were pilot-tested from September 8 to September 19, 2025, with 18 Saudi EFL teachers who were not included in the main sample. Participants commented on wording, response options, and completion burden. The median completion time was 12 minutes. Pilot alpha coefficients were .86 for the LALQ and .84 for the CBLTPS.

In the main sample, corrected item-total correlations ranged from .41 to .70 for the LALQ and from .39 to .68 for the CBLTPS. Overall and dimension-level alpha coefficients are reported in Table 3 and the Results section. Exploratory or confirmatory factor analysis was not claimed because 130 cases were insufficient for stable estimation of a 45-item measurement structure.

2.7 *Data Collection Procedures*

The survey opened only after participants confirmed that they had read the information sheet, met the eligibility criteria, and consented. No names, employee numbers, email addresses, or student information were requested. Demographic questions recorded gender, teaching level, experience, highest qualification, institution type, and country/context for sample description.

The electronic platform stored responses under anonymous single-use codes. Participants could leave before submission without penalty. The coordinating researcher exported the responses to an encrypted file, removed the invitation codes, and retained a deidentified analysis dataset. Data collection ran from October 6 to November 28, 2025.

2.8 *Missing Data and Data Analysis*

The electronic form permitted skipped instrument items so that participation remained voluntary. Four questionnaires with more than 20% of instrument items missing were excluded. Among the 130 retained cases, 11 of 5,850 expected item responses were missing across eight participants

(0.19%). A dimension score was calculated when at least four of its five items were completed, using the mean of the available items. No retained case was missing more than one item within a dimension, and demographic variables were complete.

IBM SPSS Statistics 29 was used for analysis. Reverse-coded items were recoded before composite scores were calculated. Means, standard deviations, 95% confidence intervals, and Cronbach's alpha addressed the descriptive research questions. The prespecified descriptive bands were low (1.00-2.33), moderate (2.34-3.67), and high (3.68-5.00).

Pearson correlation analysis addressed the fourth research question, and simple linear regression estimated the association between overall LALQ and CBLTPS scores. Dimension-level correlations were exploratory; Holm adjustment was applied, and all reported associations remained significant at $p < .001$. Scatterplots and residual plots showed no material nonlinearity or heteroscedasticity, standardised residuals remained within plus or minus 3, and no Cook's distance approached 1. All tests were two-tailed with $\alpha = .05$.

2.9 *Ethics Approval*

The study was approved by the Institutional Review Board of Al-Nahda University. Institutional gatekeepers authorised invitation distribution, but participation decisions were made independently by teachers. The information sheet described the study purpose, procedures, risks, confidentiality protections, voluntary participation, withdrawal before submission, and intended data use. Informed consent was recorded electronically before the questionnaire opened.

3. **Results**

This section reports the statistical results of the study in relation to the research questions. The findings are organised into four subsections. After recruitment and data-integrity information, the first substantive subsection presents the descriptive statistics of EFL teachers' Language Assessment Literacy (LAL). The second subsection reports descriptive statistics of classroom-based language testing practices. The third subsection presents the statistical relationship between Language Assessment Literacy and classroom-based language testing practices. The results are presented using descriptive and inferential statistics, including means, standard deviations, reliability coefficients, and correlation values. No interpretation or evaluative commentary is provided in this section.

3.1 *Recruitment, Response, and Data Integrity*

Of 168 eligible teachers invited, 142 submitted questionnaires (gross response rate = 84.5%). After excluding three repeated submissions, five ineligible respondents, and four questionnaires exceeding the missing-data threshold, 130 records were retained. The retained sample represented 77.4% of those invited and 91.5% of submitted questionnaires. Only 11 of 5,850 expected item responses were missing in retained records (0.19%); the prespecified within-dimension mean rule was applied as described in Section 2.8.

3.2 *Descriptive Statistics of Language Assessment Literacy*

The descriptive statistics were calculated in order to know the overall level and the distribution of the Language Assessment Literacy among the participating EFL teachers. The total mean score for the Language Assessment Literacy was $M = 3.62$ and $SD = 0.47$ which means that reported assessment literacy varies among the participants. The scores were distributed in a way that indicated that we had responses being distributed throughout the scale and no clustering in any extreme at either end.

The construct of the Language Assessment Literacy was also assessed on five dimensions that include: validity, reliability, test design and scoring, feedback-related score use, and ethics and fairness. The assessment validity dimension produced a mean of $M = 3.74$ ($SD = 0.51$). Reliability dimension registered an average score of $M = 3.68$ ($SD = 0.49$). The scoring dimension and the test design had a mean of $M = 3.59$ ($SD = 0.53$). The dimension associated with feedback-related scores registered a mean score of $M = 3.41$ ($SD = 0.56$). The ethics and fairness dimension had a mean score of $M = 3.69$ ($SD = 0.45$).

On the five dimensions, assessment validity and ethics and fairness were reported to have higher mean scores and low mean scores were reported on the feedback-related score use and test design and scoring respectively. The level of standard deviation values across all dimensions was between 0.45 and 0.56, which indicates moderate variability of responses.

Cronbach's alpha was used to test the internal consistency of the Language Assessment Literacy questionnaire. The overall reliability coefficient of the entire LAL scale was $\alpha = 0.88$, which represented high internal consistency. Individual dimension reliability coefficients ranged between $\alpha = 0.79$ and $\alpha = 0.84$, which are within the generally agreed standards of internal consistency in research done in applied linguistics.

Table 4: Descriptive Statistics of Language Assessment Literacy (N = 130)

Dimension	Mean (M)	Std. Deviation (SD)	Cronbach's α
Assessment Validity	3.74	0.51	0.84
Reliability	3.68	0.49	0.82
Test Design and Scoring	3.59	0.53	0.81
Feedback-Related Score Use	3.41	0.56	0.79
Ethics and Fairness	3.69	0.45	0.83
Overall LAL	3.62	0.47	0.88

Note. The Likert scale ranged from 1 (strongly disagree) to 5 (strongly agree).

3.3 Descriptive Statistics of Classroom-Based Language Testing Practices

Descriptive statistics were also used to analyse the distribution of classroom-based language testing practices that were reported by the participating EFL teachers. The classroom-based language testing practices had an overall mean score of $M = 3.57$ and a standard deviation of $SD = 0.44$, which implies that there was variation in the frequency of practising the various testing practices.

Classroom-based language testing practices were investigated on four dimensions, namely, construct alignment, test design practices, scoring consistency, and score interpretation and reporting. The construct alignment dimension resulted in the mean score of $M = 3.65$ ($SD = 0.48$). Test design practices dimension recorded a mean of $M = 3.52$ ($SD = 0.51$). The scoring consistency dimension had a mean score, $M = 3.61$ ($SD = 0.46$). The score interpretation and reporting dimension indicated a mean of $M = 3.50$ ($SD = 0.49$).

The findings established that construct alignment and scoring consistency were more commonly reported in comparison to the test design practices and score interpretation and reporting. Standard deviation values applied in the four dimensions ranged between 0.46 and 0.51 which suggested that there was moderate variation of responses among the respondents.

The frequency analysis showed that responses to the majority of the items were in the middle to high-range of the Likert scale. The low frequency counts occurred at the opposite ends of the scale, indicating that the responses were not concentrated on minimum or maximum values but spread generally through moderate and higher response categories.

The internal consistency of the Classroom-Based Language Testing Practices scale was established with the help of the Cronbach alpha. The overall reliability coefficient of the scale was $\alpha = 0.86$, which is an acceptable internal consistency. Coefficients of individual dimension reliability ranged between $\alpha = 0.78$ and $\alpha = 0.82$.

Table 5: Descriptive Statistics of Classroom-Based Language Testing Practices (N = 130)

Dimension	Mean (M)	Std. Deviation (SD)	Cronbach's α
Construct Alignment	3.65	0.48	0.82
Test Design Practices	3.52	0.51	0.80
Scoring Consistency	3.61	0.46	0.81
Score Interpretation and Reporting	3.50	0.49	0.78
Overall Testing Practices	3.57	0.44	0.86

Note. The Likert scale ranged from 1 (never) to 5 (always).

3.4 Relationship Between Language Assessment Literacy and Classroom-Based Language Testing Practices

Pearson correlation analysis showed a positive, statistically significant association between overall Language Assessment Literacy and overall classroom-based language testing practices, $r(128) = .63$, $p < .001$, 95% CI [.51, .72].

Correlation coefficients were also determined between individual dimensions of Language Assessment Literacy and classroom-based language testing practices. Assessment validity dimension showed significant positive correlation with construct alignment ($r = 0.58$, $p < .001$) and scoring consistency ($r = 0.55$, $p < .001$). Scoring consistency ($r = 0.60$, $p < .001$) and test design practices ($r = 0.49$, $p < .001$) were significantly related to the reliability dimension. Test design practices ($r = 0.62$, $p < .001$) had a significant association with test design and scoring literacy. The interpretation and reporting practices were significantly linked to feedback-related score use literacy ($r = 0.57$, $p < .001$). Construct alignment ($r = 0.46$, $p < .001$) and scoring consistency ($r = 0.50$, $p < .001$) had a significant relationship with ethics and fairness literacy. A simple linear regression further examined the association. The model was statistically significant, $F(1, 128) = 84.24$, $p < .001$, and explained 39.7% of observed practice-score variance, $R\text{-squared} = .397$, adjusted $R\text{-squared} = .392$. Overall LAL positively predicted reported practices, $B = 0.59$, $SE = 0.06$, 95% CI [0.46, 0.72], $\beta = .63$, $t = 9.18$, $p < .001$. The root mean square error was 0.34.

Table 6: Correlation and Regression Results for Language Assessment Literacy and Classroom-Based Language Testing Practices

Pearson Correlation Analysis

Variables	r	p-value
Overall LAL – Overall Testing Practices	0.63	< .001
Validity – Construct Alignment	0.58	< .001
Validity – Scoring Consistency	0.55	< .001
Reliability – Scoring Consistency	0.60	< .001
Reliability – Test Design Practices	0.49	< .001
Test Design & Scoring – Test Design	0.62	< .001
Feedback-Related Score Use – Interpretation	0.57	< .001
Ethics & Fairness – Construct Alignment	0.46	< .001
Ethics & Fairness – Scoring Consistency	0.50	< .001

Simple Linear Regression Analysis

Predictor Variable	β	R ²	F	p-value
Language Assessment Literacy	0.63	0.397	84.24	< .001

Note. Adjusted R-squared = .392; B = 0.59, SE = 0.06, 95% CI [0.46, 0.72], beta = .63, t = 9.18; RMSE = 0.34. Rounded values may not reproduce unrounded calculations exactly.

4. Discussion

4.1 Summary of Key Findings

The current research investigated the degree of Language Assessment literacy among EFL teachers, the prevalence of classroom-based language testing practices, and the correlation between the two constructs. The results show that among the teachers involved, the overall Language Assessment Literacy scale reported moderate mean scores, although the variation was spread among the individual dimensions. Assessment validity and ethics and fairness showed higher mean scores whereas the feedback-related score use and test design and scoring presented lower mean scores compared to higher ones. The findings also indicate that classroom-based language testing practices were reported with moderate frequency and construct alignment and scoring consistency had relatively higher mean values than the test design practices and score interpretation and reporting.

With regard to the 4th research question, the results suggest a statistically significant positive relationship between overall Language Assessment Literacy and overall classroom-based language testing practices. It was also found that significant correlations existed between specific LAL dimensions and related testing practices such as validity literacy and construct alignment, reliability literacy and scoring consistency, and feedback-related score use literacy and score interpretation and reporting.

Regression analysis also revealed that Language Assessment Literacy revealed a significant percentage of variance in reported classroom-based language testing practices. Collectively, the findings indicate a systematic relationship between assessment-based knowledge and principles on the part of teachers and their reported assessment behaviours in the classroom-based testing situations.

The participant profile gives these findings a specific scope. The sample included secondary and university teachers, public and private institutions, and varied experience and qualifications. It therefore reflects more than one institutional assessment culture, but it does not represent Saudi Arabia as a whole. Interpretations should refer to this mixed Riyadh-region sample rather than an undifferentiated global EFL population.

4.2 Interpretation of Findings

The observed variation across Language Assessment Literacy dimensions can be interpreted in light of assessment theory within applied

linguistics. Higher mean scores for validity and ethics and fairness may reflect the prominence of these concepts in assessment discourse and policy documents, where issues of fairness, transparency, and score legitimacy are frequently emphasised (Gupta, 2023; Mahar & Khoso, 2025). These dimensions are often framed as foundational principles of language assessment, which may explain their stronger representation in teachers' reported assessment literacy profiles.

In contrast, lower mean scores for feedback-related score use and test design and scoring may be associated with the technical and procedural complexity of these dimensions. Designing assessment tasks that accurately represent language constructs and applying scoring criteria consistently require specialised assessment expertise and familiarity with measurement principles (Phakiti & Steinhoff, 2023). Similarly, feedback-related score use involves interpreting assessment results in relation to test purpose and score meaning, which extends beyond surface-level familiarity with assessment concepts.

The significant associations between Language Assessment Literacy and classroom-based language testing practices suggest that assessment literacy is reflected in reported assessment behaviours. For example, the relationship between validity literacy and construct alignment indicates that teachers who report a stronger understanding of validity principles also report closer correspondence between targeted language constructs and assessment tasks.

Similarly, the correlation between reliability literacy and scoring consistency is consistent with theoretical considerations that knowledge about reliability influences scoring behaviour and consistency of judgement (Gupta, 2023). The results of the study underpin the conceptualisation of Language Assessment Literacy as a construct that goes beyond purely declarative knowledge to include principles which inform assessment-related decision-making in classroom-based testing scenarios.

4.3 *Comparison with Previous Studies*

The findings of the current study are broadly consistent with previous research in applied linguistics that has reported uneven profiles of Language Assessment Literacy among teachers. Studies have similarly noted stronger awareness of general assessment principles such as fairness and validity alongside weaker engagement with technical aspects of assessment design and score use (Tsagari & Rousoulioti, 2025; Shafii & Berger, 2025). The observed discrepancies between assessment knowledge dimensions and reported practices also align with earlier findings that teachers' self-reported assessment knowledge does not always translate uniformly into assessment behaviour (Koloi-Keaikitse & Traynor, 2023).

At the same time, the present findings extend prior work by empirically linking Language Assessment Literacy to classroom-based language testing practices. While previous studies have often examined LAL as an isolated construct, the current results provide evidence of systematic associations between specific literacy dimensions and corresponding testing practices. This addresses a gap identified in the literature, where the relationship between assessment literacy and reported testing behaviour has remained underexplored (Gal & Ryder, 2025).

Contextual factors may also account for areas of divergence across studies. Variations in institutional assessment traditions, pressure in examinations and assessment policy have been found to influence assessment behaviour in EFL settings (Prastikawati et al., 2025). These contextual constraints can help to regulate the practice of assessment literacy, although teachers might report similar degrees of assessment-related knowledge and principles.

Within Saudi institutions, comparatively strong validity and ethics scores may reflect the visibility of alignment, fairness, transparency, and confidentiality in institutional policies. Secondary teachers commonly work with Ministry-aligned outcomes and examination schedules, while university programmes may use course specifications and common assessments. These arrangements can make procedural fairness and construct alignment salient even when teachers have limited authority to alter assessment formats (Almossa & Alzahrani, 2022; Umer et al., 2018).

4.4 *Implications for Applied Linguistics*

This study contributes to applied linguistics by advancing understanding of assessment behaviour within classroom-based language testing contexts. Through an empirical study of the interaction between Language Assessment Literacy and reported testing practices, it can be seen that the findings further expand theoretical discourse on LAL beyond conceptual definitions to include behavioural correlates. The findings affirm the perspective that assessment literacy is a multidimensional construct that guides the design, scoring, and interpretation of language tests within local testing contexts.

The findings also contribute to ongoing discussions regarding theory–practice alignment in language assessment. The observed associations between literacy dimensions and testing practices provide empirical support for theoretical claims that assessment principles such as validity, reliability, and fairness are operationalised through assessment decisions and procedures (Phakiti & Steinhoff, 2023). From an applied linguistics perspective, this reinforces the importance of examining assessment as situated practice rather than abstract knowledge alone.

Besides, the research expands knowledge on classroom-based language testing as a valid site of language measurement. The findings demonstrate the relevance of the applied linguistics research to locally evolved assessments that are central in EFL situations by viewing classrooms as testing sites and teachers as evaluators. This view is in tandem with the current understanding of language testing that includes both large-scale and classroom-based testing practices.

For the participating Saudi institutions, professional development should focus on products teachers already create. Secondary-school teams could develop test blueprints linking outcomes, content, task types, and score weights. University programmes could use moderation sessions in which instructors apply rubrics to common speaking or writing samples and compare decisions. Across both sectors, teachers could analyse anonymised item statistics and student-response patterns before planning feedback and reteaching. These activities directly address the lower test-design and feedback-related score-use results.

4.5 *Limitations of the Study*

Five limitations define the evidential boundaries. First, convenience sampling from one Saudi region prevents national generalisation. Second, both instruments measured perceived literacy and reported frequency rather than demonstrated knowledge or observed practice. Third, collecting both measures in one survey increases common-method and social-desirability risk. Fourth, the cross-sectional design cannot establish direction or causality. Fifth, although content review, pilot evidence, item-total correlations, and internal consistency support initial use, the 45-item factor structure was not independently tested. Future validation should use larger, separate samples for exploratory and confirmatory factor analysis and tests of measurement invariance across teaching level and institution type.

Conclusion

The purpose of this study was to examine EFL teachers' Language Assessment Literacy, as well as its connection to classroom practices of language testing among secondary and tertiary EFL teachers in the Riyadh Region of Saudi Arabia. The investigation centred around assessment as the language measurement, the interpretation of scores and the making of decisions based on assessment, where teachers were conceptualised as assessors and classrooms as a local testing environment. Through this framing, the study was intended to contribute to the empirical research in language assessment as opposed to the discourse of pedagogical or educational evaluation.

The results suggest that the teachers who participated in the study reported moderate levels of Language Assessment Literacy on the overall scale, with significant deviation on the individual dimensions. Assessment validity and ethics and fairness were found to have higher mean scores, and comparatively lower mean scores were reported for feedback-related score use, and test design and scoring. These tendencies indicate differential engagement with the conceptual and procedural components of language assessment literacy. Simultaneously, the descriptive findings of the classroom-based language testing practice demonstrated moderate levels of reported frequencies across dimensions, with the construct alignment and scoring consistency having a higher mean value than test design practices and score interpretation and reporting.

Notably, the findings imply that there is a statistically significant positive relationship between overall Language Assessment Literacy and overall classroom-based language testing practices. Important associations were found between particular literacy dimensions and the corresponding assessment practices, such as the association linking validity literacy with construct alignment and reliability literacy with scoring consistency. Regression analysis also revealed that Language Assessment Literacy contributed a significant percentage of variation to the process of the reported testing practices, which supported the fact that assessment literacy was relevant in the context of assessment behaviour with classroom-based testing.

The most defensible conclusion is that perceived assessment competence covaried with reported assessment behaviour among teachers in the ten sampled institutions. The study does not demonstrate causal effects or verify the technical quality of teachers' actual assessment materials.

In applied linguistics, the research offers a contribution to the current debates about the nature and the extent of Language Assessment Literacy. The study expands on the previous research that has tended to study these constructs separately by empirically connecting assessment literacy dimensions with the reported testing practices. The results confirm conceptualisations of Language Assessment Literacy as a multidimensional construct which includes the principles that are reflected in assessment decisions, scoring behaviour and interpretation of assessment results.

Altogether, the research contributes to the development of the concept of classroom-based language testing as a valid area of language measurement within applied linguistics. The placement of assessment literacy in the context of local testing practices will help the study add to a more thorough reflection of how the context of language assessment theory is reflected in the assessment behaviour across EFL settings.

Directions for Future Research

Future studies should examine Language Assessment Literacy and classroom-based language testing practices across several Saudi regions and analyse secondary and tertiary sectors separately. Probability-based or stratified sampling would strengthen representativeness, while comparisons across public and private institutions could clarify how ownership, assessment policy, and moderation structures shape the translation of assessment literacy into practice.

Mixed-methods designs that combine quantitative measures of assessment literacy and qualitative analysis of assessment artefacts, including test papers, scoring rubrics, and marked scripts, could be used in future studies, as well. Such an approach may enable more detailed analysis of the relationship between reported assessment literacy and observable assessment decisions and procedures in classroom-based testing settings.

The longitudinal research designs may also be used to study stability and change in Language Assessment Literacy and assessment practices over time. Monitoring assessment behaviour through several testing cycles could inform us about the way assessment principles are maintained or remodelled in local testing environments. Also, future research could focus on examining associations between Language Assessment Literacy and other assessment-related variables, including test purpose specification or score use practices. These prescriptions may collectively expand the research on applied linguistics on language assessment behaviour within the EFL context.

Conflict of Interest: The author reported no conflict of interest.

Data Availability: All data are included in the content of the paper.

Funding Statement: The author did not obtain any funding for this research.

Declaration for Human Participants:

Ethical approval was granted by the Institutional Review Board of Al-Nahda University. Institutional permissions were obtained before recruitment. All participants provided electronic informed consent, and participation was voluntary and confidential.

References:

1. Almossa, S. Y., & Alzahrani, S. M. (2022). Assessment approaches of English language teachers in the Saudi higher education context. *Language Testing in Asia*, 12, Article 10. <https://doi.org/10.1186/s40468-022-00160-x>

2. Bravo, M. M., & Pacey, R. D. (2025). *Classroom-based assessment practices and challenges of primary teachers: Basis for a proposed upskilling program*. *Randwick International of Education and Linguistics Science Journal*, 6(2), 553–573. https://www.researchgate.net/publication/393284359_Classroom-Based_Assessment_Practices_and_Challenges_of_Primary_Teachers_Basis_for_A_Proposed_Upskilling_Program
3. Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
4. Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, 9(2), 113–132. <https://doi.org/10.1080/15434303.2011.642041>
5. Gareis, C. R., & Grant, L. W. (2015). Assessment Literacy for Teacher Candidates: A Focused Approach. *Teacher Educators' Journal*, 2015, 4–21. <https://files.eric.ed.gov/fulltext/EJ1085699.pdf>
6. Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-session language courses. *Assessment in Education*, 14(1), 75–97. <https://englishforbeachpurposes.wordpress.com/wp-content/uploads/2013/12/green-2013.pdf>
7. Gupta, K. (2023). Validity and reliability of students' assessment: case for recognition as a unified concept of valid reliability. *International Journal of Applied and Basic Medical Research*, 13(3), 129–132. <https://pmc.ncbi.nlm.nih.gov/articles/PMC10666833/#sec1>
8. Koloi-Keaikitse, S., & Traynor, A. (2023). Discrepancies between teachers' self-reported perceived skills and use of classroom assessment practices: The case of Botswana. *Cogent Education*, 10(2), 2234065. https://www.researchgate.net/publication/372291463_Discrepancies_between_teachers'_self-reported_perceived_skills_and_use_of_classroom_assessment_practices_The_case_of_Botswana
9. 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>
10. Levi, T., & Inbar-Lourie, O. (2020). Assessment literacy or language assessment literacy: Learning from the teachers. *Language*

- Assessment Quarterly, 17(2), 168-182.
<https://doi.org/10.1080/15434303.2019.1692347>
11. Mahar, H. A., & Khoso, A. R. (2025). VALIDITY IN LANGUAGE ASSESSMENT: ENSURING AUTHENTICITY, FAIRNESS, AND EFFECTIVE LEARNING OUTCOMES. *Journal of Applied Linguistics and TESOL (JALT)*, 8(4), 959-968.
https://www.researchgate.net/publication/399140779_VALIDITY_IN_LANGUAGE_ASSESSMENT_ENSURING_AUTHENTICITY_FAIRNESS_AND_EFFECTIVE_LEARNING_OUTCOMES
 12. Phakiti, A., & Steinhoff, A. (2023). Five Golden Rules for Successful Classroom Assessment Based on What We Have Learnt from JD Brown. *Language Teaching Research Quarterly*, 37, 76-90.
<https://files.eric.ed.gov/fulltext/EJ1409155.pdf>
 13. Prastikawati, E. F., Curle, S., & Vivekanantharasa, R. (2025). *Assessment in the crosshairs: How contextual factors shape English teachers' conceptions of assessment*. *Proceedings of English Language Teaching, Literature, and Translation*, 14, 719-730.
https://www.researchgate.net/publication/398031529_Assessment_in_the_Crosshairs_How_Contextual_Factors_Shape_English_Teachers'_Conceptions_of_Assessment
 14. Shafii, R. A., & Berger, J. L. (2025). Teacher assessment literacy, formative assessment practices, and their perceived efficacy in Tanzania: A scoping review. *Studies in Educational Evaluation*, 86, 101496.
<https://www.sciencedirect.com/science/article/pii/S0191491X25000537>
 15. Sheikh, M. A., Bajwa, F., Aslam, M. J., & Mir, B. (2024). Survey research: Descriptive and analytical power, types and usages. *International Journal of Politics & Social Sciences Review (IJPSSR)*, 3(III), 231-239.
https://www.researchgate.net/publication/385828054_Survey_Research_Descriptive_and_Analytical_Power_Types_and_Usages
 16. Sujianti, N. P. I. P., Sanjiyasa, I. B. M., Mesa, J., Saingo, M., & Bili, P. (2025). EVALUATING LEARNING OUTCOMES STRATEGIES FOR ENHANCING TEACHING EFFECTIVENESS IN THE CLASSROOM. *International Journal of Education and Social Science Studies*, 1(2), 69-76.
https://www.researchgate.net/publication/393590341_EVALUATING_LEARNING_OUTCOMES_STRATEGIES_FOR_ENHANCING_TEACHING_EFFECTIVENESS_IN_THE_CLASSROOM

17. Tewachew, A. (2025). EFL Teachers' Conceptions of Designing Classroom-Based Tests. *GIST–Education and Learning Research Journal*, 30(30). <https://files.eric.ed.gov/fulltext/EJ1479400.pdf>
18. Tsagari, D., & Rousoulioti, T. (2025). Enhancing assessment literacy for teachers of less commonly taught languages: Insights from greek as a second language. *Trends in Higher Education*, 4(1), 12. <https://www.mdpi.com/2813-4346/4/1/12>
19. Umer, M., Zakaria, M., & Alshara, M. (2018). Investigating Saudi university EFL teachers' assessment literacy: Theory and practice. *International Journal of English Linguistics*, 8(3), 345-356. <https://doi.org/10.5539/ijel.v8n3p345>
20. 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>
21. 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>

Appendix A. Complete Questionnaire Items

The items below reproduce the two questionnaires administered in the study. Item codes correspond to the dimensions and scoring procedures described in the Method section. Reverse-scored items are marked with the suffix R.

A1. Language Assessment Literacy Questionnaire

Prompt: Indicate how strongly you agree that each statement describes your current assessment knowledge or ability. Response options: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, and 5 = strongly agree. Items R5 and F5 are reverse-scored.

Dimension	Code	Item
Validity	V1	I can define the language ability that a classroom assessment is intended to measure.
Validity	V2	I can judge whether assessment tasks adequately represent the stated learning outcomes.
Validity	V3	I can identify task features that introduce irrelevant difficulty into a language assessment.
Validity	V4	I can distinguish evidence about score consistency from evidence supporting score interpretation and use.
Validity	V5	I can evaluate whether a proposed decision is justified by the available assessment evidence.
Reliability	R1	I can identify likely sources of inconsistency across tasks, occasions, testing conditions, or raters.
Reliability	R2	I can select procedures that improve consistency without narrowing the intended language construct.
Reliability	R3	I can explain why sampling several tasks can produce more dependable evidence than using one task.
Reliability	R4	I can evaluate whether different raters apply scoring criteria consistently.
Reliability	R5-R	Small differences between two scores always represent real differences in language ability.
Test design and scoring	T1	I can prepare a test blueprint that links outcomes, content, task types, and score weights.
Test design and scoring	T2	I can write items and tasks at an appropriate difficulty level for the intended learners.
Test design and scoring	T3	I can produce instructions that minimise misunderstanding unrelated to the target language ability.
Test design and scoring	T4	I can develop analytic or holistic scoring criteria appropriate to the assessment purpose.
Test design and scoring	T5	I can use post-assessment evidence to identify items or scoring criteria that need revision.
Feedback-related score use	F1	I can interpret patterns in results to identify learning needs rather than relying only on total scores.
Feedback-related score use	F2	I can distinguish score uses that support learning from uses intended to certify achievement.
Feedback-related score use	F3	I can explain the uncertainty and limitations that accompany classroom assessment scores.

Feedback-related score use	F4	I can combine assessment scores with other relevant evidence when making instructional decisions.
Feedback-related score use	F5-R	A total assessment score alone provides enough information for detailed feedback and reteaching.
Ethics and fairness	E1	I can identify content or task features that may unfairly disadvantage a group of learners.
Ethics and fairness	E2	I can judge whether an accommodation preserves the language construct being assessed.
Ethics and fairness	E3	I understand how to protect the confidentiality of learner scores and assessment materials.
Ethics and fairness	E4	I can consider intended and unintended consequences when choosing an assessment procedure.
Ethics and fairness	E5	I can communicate assessment criteria and decision rules transparently to learners.

Note. Dimension scores are the means of five items. The overall LALQ score is the mean of all 25 items after reverse scoring R5 and F5.

A2. Classroom-Based Language Testing Practices Scale

Prompt: Indicate how often you use each practice in your current teaching. Response options: 1 = never, 2 = rarely, 3 = sometimes, 4 = often, and 5 = always. Item SC5 is reverse-scored.

Dimension	Code	Item
Construct alignment	CA1	Before writing a test, I map each task to a stated course or curriculum learning outcome.
Construct alignment	CA2	I balance assessment tasks across the language skills and abilities named in the assessment purpose.
Construct alignment	CA3	I sample enough course content to support the intended interpretation of the score.
Construct alignment	CA4	I choose task types that match the decision the assessment is intended to support.
Construct alignment	CA5	I review the completed assessment to confirm that score weight reflects instructional priorities.
Test design practices	TD1	I prepare or consult a blueprint before finalising a classroom test.
Test design practices	TD2	I review items for ambiguity, unintended clues, and avoidable language difficulty.
Test design practices	TD3	I ask a colleague to review new high-weight items, tasks, or scoring criteria before use.
Test design practices	TD4	I use a suitable range of task formats, including meaningful language production when appropriate.
Test design practices	TD5	After administration, I examine responses to identify items or tasks that require revision.
Scoring consistency	SC1	I use written scoring criteria or rubrics when judging speaking or writing performance.
Scoring consistency	SC2	I consult sample responses or performance anchors before scoring complex tasks.
Scoring consistency	SC3	I use moderation, double-marking, or a second review when the importance of a decision warrants it.
Scoring consistency	SC4	I apply comparable administration and scoring

		procedures across learners.
Scoring consistency	SC5-R	I change scoring criteria after seeing how well or poorly the class performed.
Interpretation and reporting	SI1	I examine skill, task, or criterion-level patterns instead of reporting only a total score.
Interpretation and reporting	SI2	I provide feedback that identifies a specific strength, need, and feasible next step.
Interpretation and reporting	SI3	I explain scores in relation to the assessed learning outcomes and criteria.
Interpretation and reporting	SI4	I use assessment results to adjust subsequent teaching, practice, or support.
Interpretation and reporting	SI5	I store and communicate assessment results in ways that protect learner confidentiality.

Note. Dimension scores are the means of five items. The overall CBLTPS score is the mean of all 20 items after reverse scoring SC5.

Appendix B. Recruitment and Data-Processing Record

Recruitment/data stage	n	Disposition or calculation
Eligible teachers invited	168	Invitation denominator
Questionnaires submitted	142	Gross response rate = 84.5%
Repeated submissions	3	Excluded using single-use anonymous codes
Ineligible submissions	5	Not currently teaching EFL or not responsible for assessment
More than 20% instrument data missing	4	Excluded under prespecified rule
Final analysis sample	130	77.4% of invited; 91.5% of submitted
Missing responses in retained records	11	0.19% of 5,850 expected item responses

Note. This appendix summarises recruitment, exclusions, retention, and missing-data processing.