Exploring Assessment Literacy of Undergraduate English Language Teachers in Bangladesh: Practices and Challenges

Md Taqi Yasir

MSc in Educational Assessment, 2025

DECLARATION BY THE CANDIDATE AS AUTHOR OF THE DISSERTATION

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.
2. I allow the Department to deposit on my behalf a copy of this dissertation in the Oxford University Research Archive ('ORA') where it shall be freely available online for use in accordance with ORA's Terms and Conditions of Use [https://ora.ox.ac.uk/terms_of_use].
3. I understand that this dissertation should not contain material that can be used to personally identify individuals or specific groups of individuals (unless permission has been obtained from the individuals) and that such material should be removed before this dissertation is deposited in ORA.
4. I agree to be bound by the terms of the ORA Grant of Non-exclusive Licence [https://ora.ox.ac.uk/deposit_agreements] and I warrant that to the best of my knowledge, making my thesis available on the internet will not infringe copyright or any other rights of any other person or party, nor contain defamatory material.
5. I agree that my dissertation shall be available for download in ORA in accordance with paragraphs 2, 3 and 4 above.

Signed:	Md Taqi Yasir
Date:	2/27/2026

Acknowledgement

I would like to thank my supervisor Dr Michelle Meadows for supervising this dissertation. Her guidance helped me shape the study and make clear decisions at each stage. Her feedback made the writing stronger and more focused. I am grateful for her support throughout the process.

I also thank my parents for supporting me. Their prayers and sacrifices have brought me to this point. I thank my wife for standing by me during this long and demanding journey. She supported me every day and helped me keep going. I could not have completed this dissertation without them.

Abstract

Assessment literacy, the knowledge, skills, and principles required to design, implement, interpret, and infer educational assessments effectively, represents a critical but understudied dimension of teaching quality in higher education contexts outside North America and Western Europe. This study addresses a significant empirical gap by investigating the assessment literacy of undergraduate English language teachers in Bangladesh, an examination-dominated context with limited assessment training infrastructure. Using a cross-sectional mixed-methods design, 52 English language teachers from public and private universities completed an adapted Assessment Practices Inventory measuring self-perceived skill and self-reported frequency of use across seven assessment domains, alongside open-ended questions about challenges and professional development needs. Quantitative analysis employed non-parametric Wilcoxon signed-rank tests with Benjamini-Hochberg false discovery rate correction; qualitative data underwent inductive content analysis. Key findings revealed consistent patterns where teachers reported higher self-perceived skill ($M = 3.71$) than frequency of use ($M = 3.44$) across all domains ($p < .008$). Domain-specific analysis showed largest skill-use gaps in communication of results (gap = 0.38), construction of traditional tests (gap = 0.33), and performance-based assessment (gap = 0.29). Qualitative thematic analysis identified five explanatory themes: contextual constraints (time, class size, policy mandates), student factors (preparedness, academic integrity), lack of training and institutional support, tension between traditional examinations and alternative assessment, and technical-digital skills gaps. Teaching experience correlated positively with assessment literacy (Spearman's $\rho = .28–.40$), and formal training showed modest but consistent advantages (Cliff's $\delta = −0.28$ to $−0.34$, all non-significant after correction). No significant differences emerged between public and private universities,

suggesting uniform structural constraints across institutional types. Findings support situated models of assessment literacy, indicating that observed skill-use gaps reflect systemic barriers rather than knowledge deficits. The study demonstrates that assessment literacy is culturally specific and context-dependent, with implications for professional development and institutional reform in examination-dominated higher education systems.

Keywords: *assessment literacy; educational assessment; teacher assessment practices; skill-use gap; formative assessment; performance-based assessment; higher education; language teaching; Bangladesh; examination culture; professional development; mixed-methods research; situated assessment literacy; qualitative content analysis*.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Assessment literacy refers to the knowledge, skills, and principles that teachers require to design, implement, interpret, and infer educational assessments effectively in their classrooms and institutions. In the context of university education, assessment represents one of the most powerful influences on how students learn and what they come to value as learners. Yet despite its importance, many university teachers worldwide report limited formal training in assessment design, and substantial gaps often exist between what teachers know they should do in assessment and what they actually do in daily practice.

The problem of assessment literacy is particularly acute in higher education contexts outside North America and Western Europe. Most research on teacher assessment practices and professional development has been conducted in English-speaking countries with well-developed educational infrastructure. In contrast, university contexts in Asia, Africa, and other regions with examination-dominated systems remain significantly understudied. This gap in research matters because educational systems vary dramatically in their assessment cultures, resource constraints, infrastructure, and policy expectations. What counts as good assessment literacy looks different in a system where high-stakes written examinations drive institutional decisions than in a system designed around formative feedback and continuous learning. A framework for assessment literacy developed for American primary teachers may not fit the realities of university lecturers teaching large classes in examination-focused contexts.

Bangladesh represents such a context. Higher education in Bangladesh operates within an examination-centric culture where university admissions, degree progression, and student certification depend heavily on midterm and final written examinations. Yet little is known about how assessment literacy actually functions in this environment. To date, no published research has

quantitatively measured the assessment practices and self-perceived competence of university teachers in Bangladesh. This absence of empirical data leaves educators, administrators, and policymakers without baseline evidence to guide professional development decisions or institutional reform.

This study addresses that gap. It investigates the assessment literacy of undergraduate English language teachers in Bangladeshi universities through a mixed-methods survey design. English language teaching was selected as the focus because language assessment involves unique challenges. Assessing communicative competence, speaking proficiency, and writing quality requires judgment and skill different from assessing discrete factual knowledge. Understanding assessment literacy in language teaching contexts thus contributes both to the broader field and to a profession with specific needs.

The study employs the Assessment Practices Inventory (API), a well-validated instrument that captures both what teachers believe they can do in assessment (self-perceived skill) and what they report actually doing in practice (frequency of use). This dual-scale approach reveals the general level of assessment knowledge among teachers as well as patterns where skill exceeds practice or vice versa. Such patterns point toward barriers and constraints shaping real-world assessment decisions.

The research questions focus on describing teachers current assessment practices across multiple domains, identifying discrepancies between skill and use, and exploring whether factors such as teaching experience, formal training, and institutional type relate to assessment literacy. Qualitative data from open-ended questions illuminate the contextual, institutional, experiential, and personal factors that enable or constrain assessment practice.

Understanding assessment literacy in the Bangladeshi higher education context is essential for at

least three reasons. First, improving assessment directly improves student learning and fairness in educational decision-making. Second, this study provides the first empirical baseline in an underresearched context, enabling future comparison and evaluation of reforms. Third, the study tests whether international assessment literacy frameworks remain culturally valid in non-Western examination-dominated systems or whether they require fundamental adaptation. These questions matter not only for Bangladesh but for the growing number of university systems worldwide operating under similar constraints and examination cultures.

Chapter 2: Literature Review

Assessment literacy (AL) broadly refers to the knowledge, skills, and principles needed to design, implement, and interpret educational assessments effectively (Stiggins, 1991). This chapter traces the evolution of AL from its technical origins in the 1990s to contemporary situated frameworks, evaluates how it has been measured, reviews empirical findings on teacher AL in practice, examines critical evidence of cross cultural variation in assessment conceptions, and identifies gaps in the Bangladeshi higher education context. The chapter proceeds chronologically through four periods: early foundations (1990s), contemporary expansions (2000s to 2010s), situated and language specific frameworks (2010s to present), cross cultural factor invariance and measurement challenges, and the Bangladeshi context. It concludes with research questions that address identified gaps.

2.1 Early Foundations of Assessment Literacy (1990s)

The term "assessment literacy" emerged in the early 1990s amid growing concerns about educators' capacity to use assessment data appropriately. Stiggins (1991, p. 534) warned that society had "come to care very much about high standards of achievement" yet remained "incapable of understanding whether those standards are being met" due to widespread assessment illiteracy. He defined an assessment literate individual as one who possesses "a basic understanding of the meaning of high and low quality assessment" and the ability "to apply that knowledge to various measures of student achievement" (Stiggins, 1991, p. 536).

Early conceptions emphasised technical competence. Stiggins (1995) outlined five standards: clarity of purpose, clear targets, appropriate methods, adequate sampling, and awareness of bias. These standards reflected an assumption that technically sound tests would yield sound

educational decisions. The Standards for Teacher Competence (AFT, NCME, & NEA, 1990) formalised this view with seven core competencies including test design, scoring, grading, and ethical practices.

Plake et al. (1993) developed the Teacher Assessment Literacy Questionnaire (TALQ) based on the 1990 standards, creating the first standardised measure. However, subsequent reviews found that many such instruments lacked psychometric rigour and failed to capture emerging priorities in assessment (Gotch & French, 2014). More critically, this psychometric focus rested on an implicit assumption that assessment conceptions and practices would transfer universally across contexts, an assumption later research would fundamentally challenge.

2.2 Contemporary Expansions: Assessment for Learning (AfL) and Broadened Competencies (2000s to 2010s)

By the 2000s, scholars recognised that technical knowledge alone was insufficient. Black and Wiliam's (1998) influential review demonstrated that formative assessment strategies including providing feedback, adjusting teaching, involving students in self-assessment substantially improved student achievement. This evidence shifted the field toward "assessment for learning" (AfL) rather than solely "assessment of learning." Brookhart (2011) argued that the 1990 standards had become dated, failing to address formative practices, feedback literacy, or the social and ethical dimensions of classroom assessment. The distinction between assessment for learning (formative purposes aimed at improving ongoing instruction) and assessment of learning (summative purposes documenting achievement at a point in time) became central to contemporary frameworks, representing fundamentally different paradigms about what assessment should accomplish and what skills teachers require. In classrooms where teachers use formative assessment strategies, providing feedback, adjusting teaching, and involving students

in self-assessment, students tend to learn more and stay engaged, as multiple studies have affirmed (e.g., Harlen, 2012; Hattie, 2008; Brookhart, 2008). Conversely, poor assessment practices (e.g., unstandardised grading, unclear criteria, and lack of feedback) can demotivate learners or skew their opportunities. Thus, a teacher's assessment literacy (or lack thereof) directly affects fairness and effectiveness in the classroom. Teachers serve as 'key agents in educational assessment', and their competence in this area is integral to student learning and equity (Menken et al., 2014; DeLuca et al., 2019). A highly assessment-literate teacher is better equipped to design assessments that truly measure valued learning outcomes, to interpret results accurately, to provide constructive feedback, and to use assessment data to support each student's progress. In short, improving teachers' assessment literacy was then seen as essential for improving teaching quality and student achievement on a broad scale (Stiggins, 1999; Popham, 2011).

In response, the concept of AL evolved to encompass a broader, more dynamic set of competencies. DeLuca et al. (2016) conducted a comprehensive analysis of 15 international assessment standards documents published between 1990 and 2015. They identified eight recurring themes that define contemporary teacher AL: assessment purposes, assessment processes (design, administration, scoring, interpretation), communication of results, assessment fairness, assessment ethics, measurement theory, assessment for learning, and assessment education and support for teachers. Notably, themes such as assessment for learning and teacher professional development were virtually absent in the 1990 standards but became prominent by the 2010s. The Joint Committee on Standards for Educational Evaluation (2015) published revised Classroom Assessment Standards that explicitly incorporated formative practices, student involvement in assessment, and attention to ethical and cultural contexts.

Despite these conceptual advances, existing measurement instruments lagged behind. DeLuca et al. (2016) analysed eight prominent teacher AL instruments published between 1993 and 2012, finding that "assessment processes" (technical skills of test design and scoring) dominated item content, often representing over 50% of items. In contrast, assessment for learning, assessment fairness, equity, and teacher professional development were sparsely represented or entirely absent. Gotch and French (2014) reviewed 36 AL measures and found weak psychometric evidence for many, including low internal consistency, unstable scores as well as unclear links to student outcomes. This misalignment between what contemporary standards value and what existing tools measure indicated a critical gap. Instruments were still rooted in 1990s psychometrics while practice had moved toward situated, formative with researches on ethical dimensions.

Empirical studies using these instruments revealed persistent gaps. Studies employing the TALQ (Plake et al., 1993) and Classroom Assessment Literacy Inventory (Mertler, 2004) found that teachers could answer basic questions about matching test formats to objectives but struggled with technical aspects such as interpreting standard error or analysing item statistics (Zhang & Burry Stock, 2003). MacLellan's (2004) UK study found preservice teachers possessed only superficial conceptions of assessment and struggled to integrate it with instruction. In North America, practising teachers commonly relied on intuition rather than design principles, often mixing achievement with behaviour in grading (Brookhart, 1999; Guskey, 2003).

2.3 Situated Frameworks and Language Assessment Literacy (2010s to Present)

Recognising that AL is shaped by context, Xu and Brown (2016) proposed the Teacher Assessment Literacy in Practice (TALiP) framework reconceptualising AL as an iterative system evolving throughout careers. Teachers make compromises balancing assessment conceptions

against institutional constraints, producing negotiated settlements rather than ideal implementations. DeLuca et al. (2019) further demonstrated that AL is "negotiated, situated, and differential" across teachers and contexts in a study of 453 new teachers. The implication is profound. Assessment literacy cannot be understood as a decontextualised, transferable skillset but must be examined within the specific educational cultures and policy environments along with stakeholder relationships in which teachers operate.

2.3.1 Language Assessment Literacy as a Specialised Domain

Educators in language teaching had long recognised that assessing language proficiency presents unique challenges that generic assessment training may not address (Inbar-Lourie, 2008). Language Assessment Literacy (LAL) extends general AL by focusing on the distinctive nature of language testing. Language assessments often involve complex, subjective judgements. It includes scoring speaking or writing performance and assessing communicative competence. Henceforth, language assessments carry implications for students' identities and opportunities (Shohamy, 2001). As Taylor (2013) noted, LAL is multifaceted and varies across stakeholder profiles where classroom language teachers require different competencies than professional test developers or policymakers.

Giraldo (2018), synthesising LAL literature, proposed a framework organised into three categories. They are knowledge of assessment concepts and contexts, skills in assessment design and implementation, and principles and dispositions. Knowledge includes measurement fundamentals (reliability, validity, item analysis), language testing theory, understanding of language proficiency constructs, and awareness of local assessment contexts and policies (Inbar-Lourie, 2013; Scarino, 2013). Skills encompass test design abilities (writing quality items for all language domains, developing rubrics), classroom implementation skills (administering

assessments reliably), interpretation and feedback skills (scoring consistently, communicating results pedagogically), basic statistical literacy, and technological competence for digital assessment platforms (Davies, 2008; Fulcher, 2012). Principles include ethical responsibility, fairness, critical reflexivity about assessment's sociocultural impacts, transparency in criteria, and involving students in the assessment process (Fulcher, 2012; Scarino, 2013).

Recent LAL scholarship emphasises communities of practice and situated learning. Baker and Taylor (2024a, 2024b) advocate a collaborative, context aware approach, arguing that LAL development requires stakeholders such as teachers, test developers, administrators, policymakers in order to work together in professional networks. Taylor (2024), reflecting on 35 years in language testing, observed that effective LAL must be customised for specific communities and contexts.

2.3.2 Empirical Findings on Language Teacher Assessment Literacy

Empirical studies revealed persistent gaps in language teachers' AL, particularly in complex domains. Vogt and Tsagari (2014) surveyed foreign language teachers across several European countries, finding that while teachers valued assessment and were comfortable with informal techniques like quizzes, many lacked training in complex tasks such as developing tests with clear specifications or conducting item analysis. A significant number reported never having formal coursework on language assessment. Gan et al. (2018) found similar patterns among English teachers in China, who felt uncertain about designing performance based assessments and using scoring rubrics due to minimal training opportunities. When teachers lacked confidence, they might avoid certain assessment types. For example, sticking to multiple choice tests and avoiding speaking assessments because scoring is perceived as difficult which potentially narrowed the assessment of students' abilities and undermined validity (Crusan et al.,

2016).

Common challenges across contexts included lack of time to design or refine assessments, large class sizes constraining personalised feedback, external examinations driving teaching priorities and encouraging teaching to the test, and limited access to mentoring and structured professional support (Vogt & Tsagari, 2014; Tsagari, 2021; DeLuca & Bellara, 2013). These structural constraints indicated that even teachers with strong conceptual knowledge might struggle to enact practices due to contextual barriers rather than individual incompetence (Xu & Brown, 2016). This insight is critical. Low reported use of sophisticated assessment practices may not reflect knowledge deficits but rather systemic constraints that prevent implementation.

2.3.3 Methodological Insights from Empirical Studies

A critical gap in the literature is the limited presentation of empirical methodologies. Most AL research relies on self-report surveys or knowledge quizzes, which may not fully reveal practical assessment ability. For example, a teacher might know the definition of "reliability" and answer correctly on a quiz but still design classroom tests that yield unreliable results (Gotch & French, 2014). Zhang and Burry Stock (2003) advanced the field by introducing a dual scale format in the Assessment Practices Inventory (API), asking teachers to rate both their skill level and frequency of use for each assessment practice. This design, employed in surveys across diverse contexts including China (Gan et al., 2018), enabled identification of patterns where teachers felt capable but reported low enactment. This suggested structural barriers rather than knowledge deficits. However, DeLuca et al. (2019) cautioned that self-report data are subject to social desirability bias, limited self-awareness, and recall limitations, indicating the need for mixed methods designs that combine surveys with qualitative interviews or classroom observations to triangulate findings and reveal the gap between espoused beliefs and applied practices.

2.4 Cross Cultural Factor Invariance and the Challenge of Universality

A critical but often overlooked issue in assessment literacy research concerns whether the construct itself operates equivalently across cultural and educational contexts. There had been systematic investigation of this question through multi group confirmatory factor analysis studies examining teacher conceptions of assessment across diverse jurisdictions. The findings fundamentally challenge assumptions of universality in AL frameworks and instruments.

2.4.1 Evidence of Cultural Specificity in Assessment Conceptions

Brown et al. (2019) conducted a landmark cross cultural validation study examining eight previously published models of teacher conceptions of assessment across 11 datasets from multiple countries (New Zealand, Queensland, Hong Kong, India, Cyprus, Egypt, Spain, Ecuador). Using nested multi group confirmatory factor analysis, they tested whether any published model achieved configural, metric, and scalar invariance across contexts. The results were striking: only one model (from India) achieved configural invariance across all 11 datasets, and even this model failed to achieve metric equivalence. These findings led Brown et al. (2019, p. 16) to conclude that "context, culture, and local factors shape teacher conceptions of assessment" and that "there is indeed no single global model."

This lack of invariance had profound implications. Teachers in low stakes formative assessment environments (New Zealand, Queensland, Cyprus, Catalonia) conceived of assessment differently than teachers in examination dominated systems (Hong Kong, Egypt, India, Ecuador). In New Zealand, where high stakes testing has been largely absent from primary education since the 1940s, teachers strongly endorsed assessment as improving teaching and learning while rejecting its use for accountability (Brown, 2004). In contrast, Hong Kong teachers strongly

associated assessment for improvement with assessment for making students accountable, reflecting an examination culture where these purposes are inseparable (Brown, 2004). Egyptian teachers, working within an education system where end of year exams are the sole mechanism for student progression, conceived of assessment primarily through the lens of summative evaluation and certification, despite policy rhetoric about formative assessment (Gebril & Eid, 2017).

2.4.2 Factor Structure Differences across Contexts

Brown and Remesal (2017) found that the original four factor model of teacher conceptions (improvement, irrelevance, school accountability, student accountability) was inadmissible when tested with Spanish and Ecuadorian samples due to negative error variances and positive not definite covariance matrices. Alternative models had to be constructed for each context. For Spanish teachers, assessment for improvement was weakly distinguished from assessment for accountability, reflecting a policy environment where continuous school based evaluation served both purposes simultaneously. In Ecuador, a new "caution" factor emerged, representing teachers' concerns about assessment's potential negative consequences for students, a dimension absent in Western models.

Similarly, Brown and Remesal (2012) compared prospective teachers in New Zealand and Spain, finding that while both groups responded to the same inventory items, the underlying factor structure differed significantly. Only configural invariance was achieved, meaning the pattern of factors was similar but the strength of item loadings and factor relationships varied substantially. Spanish prospective teachers gave similar mean scores across most factors (ranging only from 3.23 to 3.56 on a six point scale), whereas New Zealand prospective teachers made strong distinctions, ranging from 2.42 (assessment is bad) to 4.49 (assessment improves learning). This

suggested that Spanish respondents either held more ambivalent beliefs or were less willing to make strong evaluative distinctions, possibly reflecting different cultural norms around expressing professional opinions.

2.4.3 Implications for Assessment Literacy Frameworks and Measures

These factor invariance problems had several critical implications. First, they demonstrate that assessment literacy is not a culturally neutral construct that can be measured identically worldwide. What counts as "assessment literate" varies depending on whether the educational system privileges formative continuous assessment, high stakes examinations, teacher judgement, or external standardisation (Xu & Brown, 2016). Second, they reveal that instruments developed in one context cannot simply be translated and administered elsewhere without substantial adaptation and revalidation. The Teacher Conceptions of Assessment inventory, despite being widely used internationally, produced different factor structures in different contexts, meaning that cross cultural comparisons using raw scale scores were potentially invalid (Brown et al., 2019).

Third, and most fundamentally, these findings challenged the very notion that there is a single universal "assessment literacy" that all teachers should possess. Instead, they support a situated view in which assessment literacy must be defined relative to the assessment culture, policy expectations, and educational values of specific contexts (Xu & Brown, 2016). A teacher who is highly assessment literate in a New Zealand formative assessment environment might struggle in a Bangladeshi examination focused system, not due to lack of knowledge but due to misalignment between their assessment conceptions and the contextual demands. This insight has vast significance for professional development and for research. Both must be designed with

explicit attention to the local assessment ecology rather than assuming that generic international frameworks can be applied universally.

2.5 Assessment Literacy in Bangladesh: Critical Gaps in Higher Education English Programmes

The discussion thus far has outlined global developments in assessment literacy and revealed fundamental challenges in assuming universality. However, a notable gap exists in contexts like Bangladesh, especially in higher education English language teaching. Most AL research has been conducted in North America, Europe, and East Asia, often in primary or secondary education. In contrast, very little empirical work has examined the assessment literacy of university level English instructors in Bangladesh. This gap is significant because Bangladesh's educational landscape has unique features. Bangladesh has an examination dominated culture, resource constraints, limited assessment training scopes which may shape teachers' practices and needs in distinct ways.

2.5.1 The Examination Centric Culture

The education system in Bangladesh is highly examination oriented. Exams are the primary mode of assessment used to make decisions about student progression, certification, and placement (Sultana, 2019). University English programmes often follow this pattern, with student achievement largely determined by midterm and final examinations that emphasise written tests. Alternative forms of assessment such as oral presentations, portfolios, performance tasks are used less frequently, though some private or internationally linked universities might have started to introduce them (Das et al., 2014). The heavy weight placed on high stakes tests can encourage a "teaching to the test" culture, sometimes at the expense of formative assessment or feedback for learning (Khan, 2022).

A documented misalignment exists between communicative language teaching goals in the curriculum and the traditional forms of assessment actually used, indicating an implementation gap (Das et al., 2014). Teachers in such contexts may not feel empowered to innovate in assessment if institutional expectations and training have prepared them only to administer standardised tests and paper exams. Given the factor invariance problems documented by Brown et al. (2019), it is highly likely that Bangladeshi teachers' conceptions of assessment differ substantially from those in Western contexts, with greater emphasis on assessment as certification and accountability and potentially weaker endorsement of formative purposes, simply because that reflects the assessment reality in which they work.

2.5.2 Limited Training and Institutional Support

Teachers at the university level in Bangladesh are typically content experts in literature, linguistics, or applied linguistics, but they may not receive substantial training in test design or effective use of classroom assessments. A telling example comes from national examinations. Teachers serving as exam setters often lacked formal assessment training (Sultana, 2019, p. 89), perpetuating ritualistic practices.

Though this example is from the school sector, it indicates a broader trend. Many educators in Bangladesh, including those teaching English at the tertiary level, may have never had in depth instruction on assessment design, analysis, feedback techniques, or formative assessment practices (Khan, 2022). As a result, their assessment practices might rely on what they experienced as students or simply follow longstanding institutional routines. There is likely heavy reliance on written exams and perhaps some quizzes or assignments, but less frequent use of criterion referenced rubrics, structured feedback protocols, or performance based assessments (Crusan et al., 2016). The situated nature of assessment literacy (Xu & Brown, 2016) suggested

that without systemic changes in policy and resources, and professional development infrastructure, individual teacher knowledge gains may have limited impact on actual practice.

2.5.3 Absence of Empirical Data on Higher Education Assessment Literacy

To date, no published research in Bangladesh has quantitatively measured university English teachers' self-reported skill and use profiles across the key domains of assessment literacy. This absence is problematic because without empirical data, policymakers and educators have little evidence to identify professional development needs or craft reforms that would improve assessment practices in higher education. Understanding what Bangladeshi tertiary teachers actually do in their assessment routines (Use) and what they believe they can do (Skill) is crucial for designing relevant and effective professional development. Such insights can bridge the disconnect between international standards and local practice, ensuring that capacity building initiatives address the realities on the ground rather than importing decontextualised Western frameworks that may not fit the Bangladeshi assessment ecology. Investigating how teacher characteristics such as institution type (public versus private), teaching experience, and prior assessment training which relate to assessment literacy could help determine whether certain groups have specific needs.

2.6 Summary and Research Questions

This review traced AL from technical origins to situated frameworks, demonstrating that AL is culturally specific rather than universal (Brown et al., 2019). Critically, cross cultural researches had demonstrated that assessment literacy is not a universal construct. Factor invariance studies across 11 jurisdictions found no single global model of teacher conceptions of assessment, with substantial variation between low stakes formative systems and examination dominated contexts.

This challenges assumptions underlying many AL frameworks and instruments. It also revealed that what counts as assessment literate varies systematically with policy environment and cultural values. The lack of metric and scalar invariance means that international comparisons using standard instruments may be invalid, and that professional development must be designed with explicit attention to local assessment ecologies.

In the specific domain of language assessment literacy, unique challenges emerge around assessing communicative competence, scoring subjective performances, and operating within diverse linguistic and cultural contexts. Recent scholarship emphasised that LAL must be developed through communities of practice and customised to specific stakeholder needs and local contexts.

A critical gap exists in the Bangladeshi higher education English teaching context. Despite an examination centric culture and limited assessment training among university instructors, no empirical research has profiled their assessment literacy. This study addresses that gap by examining undergraduate English teachers' self-reported assessment skills and frequency of use across key AL domains, identifying discrepancies between self-perceived capability and self-reported practice, and exploring how institutional and experiential factors relate to these patterns. Given the cultural specificity documented in Brown et al.'s (2019) work, this study also investigates whether international AL frameworks require adaptation for the Bangladeshi context.

2.7 Research Questions

The study is therefore guided by five research questions (RQs):

RQ1: What are undergraduate English teachers perceived skills in assessment across the key domains of assessment literacy?

RQ2: How frequently do these teachers use various assessment practices across the same domains in their current teaching?

RQ3 (O3): What gaps exist between teachers perceived skill and their actual use of assessment practices by domain? Which assessment domains show the most significant discrepancies between Skill and Use?

RQ4 (O4): To what extent do factors such as institution type, teaching experience, and prior assessment training relate to teachers' assessment literacy (regarding their Skill ratings, Use frequencies, and Skill–Use gaps)?

RQ5 (O5): What challenges do teachers report in implementing assessment, and what professional development priorities do they identify for improving assessment in undergraduate English language teaching?

Each of these questions is designed to yield actionable insights. RQ1 and RQ2 establish baseline profiles of self-perceived competence and behaviour. RQ3 highlights where mismatches occur, signalling potential areas of unmet need or inefficiencies. RQ4 allows us to see if different groups of teachers require different support (for example, if less experienced teachers struggle more with certain assessment tasks). Finally, RQ5 adds a qualitative and thematic lens, giving teachers a voice to elaborate on their circumstances and suggest what would help them. The answers to these questions will provide a comprehensive picture of assessment literacy among higher education English teachers in Bangladesh. They will directly address the gaps identified in the literature review.

Chapter 3. Research Methodology

3.1 Research Design

This study adopted a cross sectional survey design that combined quantitative and qualitative approaches to investigate the assessment literacy of undergraduate English language teachers in Bangladesh. Cross sectional designs collect data from participants at a single point in time, providing a snapshot of current practices, perceptions, and relationships among variables (Creswell & Creswell, 2018). This approach is appropriate when the research aims to describe the state of a phenomenon at a particular moment rather than track changes over time (Wang & Cheng, 2020).

The cross sectional survey design was selected for several reasons. First, the research questions focused on understanding teachers' current self-perceived assessment skills, their reported frequency of using various assessment practices, and the gaps between these dimensions. A single time point measurement was sufficient to address these descriptive and correlational objectives (Rindfleisch et al., 2008). Second, cross sectional surveys are practical and cost effective for collecting data from geographically dispersed participants across multiple institutions in Bangladesh (Lavrakas, 2008). Third, this design enabled the collection of both numerical data through Likert scale ratings and qualitative insights through open ended questions, facilitating comprehensive understanding (Creswell & Plano Clark, 2018).

However, cross sectional designs have inherent limitations. Data collected at only one point in time cannot establish causal relationships between variables or track how assessment practices develop over time (Mann, 2003). The findings represent a snapshot that may be influenced by temporary contextual factors such as recent policy changes or curriculum reforms (Sedgwick,

2014). Despite these limitations, the cross sectional survey design was the most appropriate choice given the study's aims and the need to establish baseline data on assessment literacy in this underresearched context.

3.2 Participants and Sampling

3.2.1 Population and Sampling

The study targeted undergraduate English language teachers working in higher education institutions across Bangladesh. The target population included faculty members actively teaching undergraduate English language courses in public universities, private universities, and other tertiary institutions. Bangladesh's higher education sector is diverse, with variations in institutional resources, class sizes, and teaching conditions (Alam et al., 2020; Rahman & Pandian, 2018).

The study employed a combination of purposive and convenience sampling. Purposive sampling was used to ensure that participants possessed characteristics relevant to the research focus, specifically, experience teaching undergraduate English language courses. This ensured all respondents had direct experience with the assessment practices under investigation (Tongco, 2007). Convenience sampling was used within this purposive framework to facilitate practical recruitment. Participants were recruited through accessible channels, including institutional contacts, professional networks, and social media groups for English language educators in Bangladesh (Dörnyei, 2007). Existing professional connections and institutional relationships were utilised to distribute the survey link and invite participation.

Teachers were eligible to participate if they met two criteria: they must have been actively teaching undergraduate English language courses at the time of data collection, and they must

have been willing to participate voluntarily and provide informed consent. No restrictions were placed on years of teaching experience; the sample deliberately included teachers ranging from novice to veteran educators, as variation in experience was a descriptive characteristic of interest (Zhang & Burry Stock, 2003).

Recruitment occurred through English department heads, professional networks, and social media groups. All recruitment materials clearly explained the voluntary nature of participation, the time commitment required (approximately 15 to 20 minutes), and confidentiality protections. Participants were informed they could withdraw at any time before submitting responses.

A target of 50 to 100 respondents was anticipated based on practical considerations and precedents in similar research using the Assessment Practices Inventory (DeLuca et al., 2016; Zhang & Burry Stock, 2003). The final sample consisted of 52 participants which was adequate for descriptive analyses (Field, 2018).

3.3 Instrumentation

3.3.1 Rationale for Instrument Selection

Selecting an appropriate instrument to measure teacher assessment literacy is a critical decision that shapes the validity and utility of research findings (DeLuca et al., 2016). Several established instruments exist in the literature, each with distinct strengths, limitations, and theoretical orientations (Gotch & French, 2014). This study adapted the Assessment Practices Inventory (API), originally developed by Zhang and Burry Stock (1997, 2003). The decision to use the API rather than develop a new instrument or select a different existing measure was based on careful consideration of the research aims, the construct being measured, the context of the study, and the psychometric evidence available for various instruments. The Assessment Practices

Inventory (API; Zhang & Burry-Stock, 1997, 2003) was selected over alternative instruments such as the Assessment Literacy Inventory (Plake et al., 1993) or Classroom Assessment Literacy Inventory (Mertler, 2004) because it uniquely employs a dual-scale format measuring both self-perceived skill and frequency of use, which directly addresses this study's central research question about the gap between teacher's self-perceived capability and enactment.

It is critical to emphasize that the API measures self-reported perceptions of skill and frequency, not objectively assessed competence or directly observed practice (Gotch & French, 2014). Throughout this study, "skill" refers to teachers' self-perceived capability, and "use" refers to self-reported frequency of use/implementation of practices. These are subjective ratings that may be influenced by social desirability, limited self-awareness, or recall bias (Kruger & Dunning, 1999; Paulhus & Vazire, 2007). The findings therefore reflect teachers' subjective profiles of their assessment literacy rather than validated measures of actual proficiency.

The API was selected for several other reasons. First, it is one of the most widely used and well validated instruments for measuring teachers' self-reported assessment competence and practices (DeLuca et al., 2016; Gotch & French, 2014). The original API has been validated across multiple studies and contexts (Zhang & Burry Stock, 1997). This strong foundation provided confidence that the instrument could produce consistent and meaningful data.

Second, the API's dual scale format, which asks teachers to rate both their skill level and their frequency of use for each assessment practice, aligned well with the study's research questions. Although Skill and Use employ different scale anchors (competence vs. frequency) and therefore do not measure the same underlying construct, this design enables researchers to examine patterns of alignment or divergence between teachers' self-perceived capability and their self-

reported implementation. Specifically, it allows identification of practices where teachers report feeling capable but indicate less frequent use, which may reflect structural barriers, contextual constraints, or other factors affecting practice beyond individual competence (Zhang & Burry-Stock, 2003). This dual-scale approach is therefore valuable for profiling the relationship between perceived assessment literacy and reported practice, even though it does not measure an objective "gap" between actual knowledge and actual behaviour.

Third, the API covered a comprehensive range of assessment practices across multiple domains, including test construction, performance assessment, grading, communication of results, and ethical considerations (Zhang & Burry Stock, 2003). This breadth allowed the study to profile teachers' assessment literacy holistically rather than focusing narrowly on one or two aspects of assessment. The seven factor structure identified through factor analysis in the original validation study provided a theoretically grounded framework for organizing and analyzing the data (Zhang & Burry Stock, 1997).

The API is a 67-item questionnaire validated across multiple contexts (Zhang & Burry-Stock, 1997, 2003). It uses a dual-scale format asking teachers to rate both skill and frequency of use for each practice, enabling identification of capability-execution gaps.

However, it is important to acknowledge the limitations of the API and why it may not perfectly align with all aspects of the study's theoretical framework. The API is a self-report instrument, meaning it measures teachers' perceptions of their skills and self-reported frequencies of use, not their actual competence or observed classroom practices (Gotch & French, 2014). Self-report data are subject to several biases. Teachers may overestimate their skills due to social desirability bias, the tendency to present oneself in a favorable light (Paulhus & Vazire, 2007). They may

also lack accurate self-awareness of their true competence, particularly if they have not received formal training in assessment (Kruger & Dunning, 1999). Similarly, self-reported frequency of use may be influenced by memory biases or by what teachers believe they should be doing rather than what they actually do (Schaeffer & Presser, 2003).

Despite these limitations, self-report instruments remain valuable and widely used in educational research for several reasons (Kunter & Baumert, 2006). They are practical and cost effective for collecting data from large or geographically dispersed samples. They provide insight into teachers' subjective experiences, beliefs, and intentions, which influence their classroom behaviors even if they do not perfectly predict those behaviors (Ajzen, 1991). For the purposes of this study, understanding teachers' perceptions of their skills and their reported practices is valuable in its own right, particularly in a context like Bangladesh where little prior research exists. The self-report data provide a starting point for identifying areas where teachers perceive gaps between their competence and their practice, which can inform the design of professional development programs even if the data do not capture objective competence (DeLuca et al., 2016).

3.3.2 Adaptations for the Bangladesh Context

For the present study, the API was adapted to fit the context of undergraduate English language teaching in Bangladesh. Adaptations were necessary to ensure that the items were relevant, understandable, and meaningful to the target population of university English faculty. The adaptation process involved modifying item wording, adjusting examples to reflect higher education contexts, adding demographic questions, and implementing the dual scale response format. Each type of adaptation is described below.

3.3.2.1 Wording and Contextual Modifications

Many items were reworded or contextualized to better reflect Bangladeshi university assessment practices in English language programs. The original API assumed a K-12 classroom context in the United States, so terminology and references were adjusted to fit the higher education environment. For example, generic references to "students" and "classrooms" were retained where appropriate, but terms like "class" were often replaced with "course" to align with university terminology. References to "parents," which are common in school settings, were removed or changed to references to "students," "academic committees," or "colleagues," as parent involvement is not typical at the tertiary level in Bangladesh.

Items that mentioned standardized tests were revised to focus on assessments more relevant to university faculty, such as entrance exams, placement tests, or departmental exams. For instance, an item about interpreting standardized test results for instructional planning was modified to refer to interpreting placement or proficiency exam results, which university lecturers sometimes review when planning courses for incoming students.

Examples were inserted into item descriptions where necessary to make them more concrete and relatable. For instance, the phrase "performance assessment" was clarified with examples relevant to English language courses, such as oral presentations, debates, essays, and portfolio assessments. This helped ensure that respondents understood what each item was referring to, even if they were not familiar with assessment terminology from the Western educational measurement literature.

A few items from the original API that addressed practices entirely absent from Bangladeshi universities were considered for omission. For example, interpreting and using standardized

national test statistics for program evaluation is not a routine part of most university lecturers' duties in Bangladesh, where national level standardized testing is not common in higher education. However, most items were retained with modifications rather than removed, to preserve the breadth of the instrument and to maintain comparability with the original factor structure where possible.

3.3.2.2 Additional Demographic Questions

Demographic questions included teaching experience, qualifications, academic rank, institution type, and prior assessment training.

3.3.2.3 Dual Scale Response Format: Frequency and Skill

Following the precedent set by Zhang and Burry Stock (2003), the adapted API used a dual rating response format for each of the 67 assessment practice items. Each item used dual ratings: Frequency (1=Never to 5=Very Often) and Skill (1=Not at all skilled to 5=Highly skilled). This design distinguishes knowledge from enactment (Zhang & Burry-Stock, 2003).

The dual scale format has been shown to have strong reliability in previous studies. Zhang and Burry-Stock (2003) reported Cronbach's alpha values of 0.77 to 0.89 for the frequency subscales and 0.85 to 0.91 for the skill subscales.

3.3.2.4 Open Ended Questions

Two open-ended questions captured challenges and professional development needs, providing qualitative depth (Creswell & Creswell, 2018).

3.3.2.5 Presentation of Scale Anchors

The adapted survey ensured that scale anchors were presented alongside each item. Rather than only providing the numeric values (1 to 5) and expecting respondents to remember what each number meant, the online form displayed the full text of the anchors (e.g., "Never," "Rarely," "Sometimes," "Often," "Very often") as labeled response options for each item (see Appendix 1 for visualisation). This improved clarity and reduced the cognitive burden on participants, making it easier for them to provide accurate and thoughtful ratings (Dillman et al., 2014).

3.3.3 Limitations and Justification: API Coverage vs. Contemporary Frameworks

The literature review (Chapter 2) drew extensively on contemporary assessment literacy frameworks that emphasize formative assessment, feedback literacy, assessment for learning, and assessment as a situated social practice embedded in specific disciplinary and institutional contexts (Xu & Brown, 2016; Pastore & Andrade, 2019; Inbar-Lourie, 2017). These frameworks position assessment literacy as a practice-oriented, context-sensitive competence that develops through participation in communities of practice rather than as a static set of technical skills (DeLuca et al., 2016).

The API, by contrast, reflects an earlier, more traditional conceptualization of assessment literacy rooted in psychometric principles and classroom measurement (Zhang & Burry-Stock, 1997). Its items focus predominantly on test construction, grading, standardised test interpretation, and ethical considerations. While it includes some items on performance assessment, feedback, and communicating results, it does not fully capture the contemporary emphasis on assessment as learning, student self-assessment, peer assessment, or the sociocultural and institutional dimensions of assessment practice emphasised in recent language assessment literacy research (Fulcher, 2012; Inbar-Lourie, 2017).

This mismatch between the literature review's theoretical scope and the instrument's operationalisation is deliberate and pragmatically justified for three reasons. First, no existing validated instrument fully captures the contemporary, situated view of assessment literacy in a way that is practical for large-scale survey research (Pastore & Andrade, 2019). Developing and validating a new instrument was beyond the scope and resources of this study. Second, despite its traditional focus, the API covers foundational assessment practices such as test quality, performance tasks, grading, feedback, and communication that remain central to effective assessment and student learning (Black & Wiliam, 1998; Brookhart, 2013). These practices are relevant and meaningful in the Bangladeshi university context, where formal examinations dominate but where there is also growing interest in alternative assessment methods. Third, the qualitative component of the study (open-ended questions) was designed to capture aspects of teachers' experiences, challenges, and contextual constraints that the quantitative items might miss, thereby providing a more complete picture when both data sources are integrated (Creswell & Plano Clark, 2018).

3.4 Data Collection

3.4.1 Data Collection Strategy and Administration

The survey was administered online using Microsoft Forms over approximately four weeks. Online administration enabled geographically dispersed participation, facilitated data management, and supported anonymity. Participants received a participant information sheet detailing the study's purpose, procedures, voluntary nature, confidentiality protections, and ethics contact information. Informed consent was obtained before participants could access the survey. No personal identifiers were collected; responses remained completely anonymous.

3.5 Data Analysis

3.5.1 Data Preparation and Cleaning

Before conducting analysis, raw data were carefully prepared and cleaned to ensure quality and integrity. Responses completed in implausibly short amounts of time were flagged for review, as they might indicate disengagement. Straight lining patterns, where respondents selected identical responses across many items, were identified using standard deviation calculations. For example, if a participant rated all 67 items as "3" for both skill and frequency, this would suggest a lack of engagement rather than genuine reflection. Straight-lining (identical responses across items) was identified via standard deviation calculations. Respondents with SD ≈ 0 were excluded (Kim et al., 2019).

Missing data patterns were examined; cases with substantial missingness were excluded. The survey platform allowed participants to skip items if they chose, although most items were not mandatory. The amount and pattern of missing data were assessed. If missingness appeared to be random and only a small percentage of items were missing for a given respondent, the data were retained and missing values were handled using pairwise deletion in analyses (Little & Rubin, 2002). If a large proportion of items were missing for a respondent, indicating that they did not complete a substantial part of the survey, the response was excluded from analysis.

Implausible or inconsistent response patterns were reviewed. For instance, if a respondent reported very high frequency of use (e.g., "very often") for practices that are logically incompatible with one another, or if qualitative responses contradicted quantitative ratings in ways that suggested misunderstanding of the questions, the response was reviewed. However, because inconsistencies can also reflect genuine complexity or ambivalence, responses were not

automatically excluded for this reason unless there was clear evidence of random or careless responding.

After cleaning, the dataset contained 52 valid responses with complete scale level data and no out of range values. The decisions made during data cleaning were documented, including the number of responses excluded and reasons for exclusion (Van den Broeck et al., 2005).

3.5.2 Quantitative analysis

Descriptive statistics were calculated to summarize the characteristics of the sample and the distribution of responses on the assessment literacy scales. For demographic variables such as years of teaching experience, measures of central tendency (mean and median) and dispersion (standard deviation and range) were computed. Composite scores were created by taking the mean of items within each subscale, yielding scores on the same 1 to 5 scale as individual items (Nunnally & Bernstein, 1994). For categorical variables such as institution type and prior training, frequencies and percentages were reported (Field, 2018).

Internal consistency reliability was assessed for each subscale and for the overall skill and frequency scales using Cronbach's alpha coefficient. Cronbach's alpha estimates the degree to which items within a scale measure the same underlying construct, with higher values indicating greater consistency (Cronbach, 1951). Alpha values of 0.70 or higher are generally considered acceptable for research purposes, and values of 0.80 or higher are considered good (Nunnally & Bernstein, 1994).

Non parametric tests were selected for the primary inferential analyses for three reasons. First, the data are based on Likert scale ratings, which are technically ordinal rather than interval level measurements (Jamieson, 2004). Non parametric tests are appropriate for ordinal data because

they make inferences based on ranks rather than assuming equal intervals (Conover, 1999). Second, preliminary examination using Shapiro-Wilk tests revealed that several subscale scores were not normally distributed ($p < .05$), violating the normality assumption required for parametric tests (Field, 2018). Third, the sample size was relatively small ($N = 52$), limiting parametric test power. These three considerations together: the ordinal nature of Likert data, observed non-normality, and small sample size, each independently justified the use of non-parametric methods, making them the most defensible analytical choice for this study.

The Wilcoxon signed rank test (Wilcoxon, 1945) served as the non-parametric analogue of the paired samples t test, comparing teachers' self-reported Skill with their reported Use of assessment practices at the scale level across seven API domains and the overall composite. The test evaluated whether the within teacher median difference (Skill minus Use) exceeded zero (Cohen et al., 2018; Conover, 1999; Field, 2018). Because seven domain tests were conducted, the probability of false positives was managed using the Benjamini Hochberg false discovery rate (FDR) procedure at $q = 0.05$ (Benjamini & Hochberg, 1995; Benjamini, 2010). Adjusted p values below 0.05 were treated as statistically significant. Effect sizes were expressed as rank biserial r, interpreted using approximate thresholds of 0.10 (small), 0.30 (medium), and 0.50 (large) to aid practical interpretation (Cohen, 1988; Tomczak & Tomczak, 2014).

The Mann Whitney U test (Mann & Whitney, 1947; Conover, 1999) examined differences in assessment literacy by institution type (public versus private) and prior training (yes versus no). This test compares distributions of two independent groups by ranking scores, appropriate when grouping variables are independent and outcomes are ordinal (Field, 2018). Spearman's rho correlation assessed monotonic relationships between teaching experience (years) and assessment literacy scores, appropriate for ordinal outcomes and skewed distributions because it

relies on ranks rather than assuming linearity (Conover, 1999; Field, 2018). All quantitative analyses were conducted using R (v4.4.1) and SPSS (IBM Corp., 2024).

3.5.3 Qualitative analysis

Open ended responses were analyzed using inductive content analysis. Responses were imported into NVivo qualitative data analysis software (QSR International, 2024) and systematically coded to identify patterns. Content analysis is a systematic method for identifying, coding, and categorising patterns or themes in textual data (Hsieh & Shannon, 2005). An inductive approach means that themes and categories emerge from the data itself rather than being imposed a priori based on theory (Elo & Kyngäs, 2008).

Open coding labeled segments of text describing challenges or training needs. Codes were iteratively grouped into broader themes. Throughout the analysis, care was taken to maintain anonymity. The analysis maintained confidentiality by removing identifying details and referring to participants by generic identifiers (Kaiser, 2009).

Qualitative findings were integrated with quantitative results to provide richer understanding. For instance, if quantitative data showed gaps between skill and use, qualitative data could explain contextual barriers contributing to those gaps (Creswell & Plano Clark, 2018).

3.6 Ethical Considerations

The study received ethical approval from Oxford's Department of Education Research Ethics Committee (Ethics reference Education Educ DREC 1891578). Participants received detailed information sheets and provided informed consent before completing the survey. The survey was anonymous, preventing identification. Data were securely stored on password protected systems

with encryption. Participants could withdraw before submitting responses, though submitted anonymous data could not be retrieved.

Chapter 4. Results

4.1 Quantitative and Qualitative Data Analysis

This chapter presents the quantitative and qualitative findings from the adapted Assessment Practices Inventory completed by 52 undergraduate English language teachers in Bangladeshi universities. The quantitative strand reports descriptive statistics, internal consistency estimates, and non-parametric tests comparing self-reported Skill and Use scores across seven assessment domains. The qualitative strand summarises themes from open-ended responses on assessment challenges and desired professional development, providing contextual depth to help interpret the Skill–Use patterns.

4.2 Descriptive Profile of Participants and Assessment Practices

4.2.1 Participant Characteristics and Institutional Context

The sample consisted of 52 undergraduate English teachers working in public and private universities in Bangladesh. As shown in Table 1, 33 participants were based in private universities and 19 in public universities. Most respondents held a Master's degree and were employed at lecturer or assistant professor level, reflecting early- to mid-career academics rather than senior management or purely administrative staff.

Institutionally, these teachers were positioned in different ways. Many worked within English departments or language centres embedded in larger faculties, while others reported roles in more centralised language programmes serving multiple departments. These arrangements matter for assessment literacy. Teachers embedded in departments may have more autonomy over course-level assessment design, whereas those working centrally are often required to implement faculty- or university-wide assessment schemes with limited flexibility. This variation in how

teachers are situated helps to explain why some respondents described themselves as technically capable yet constrained by programme-level policies and exam regulations.

The sample was geographically concentrated in Dhaka (n = 47), with a smaller number from Khulna, Chittagong, Tangail, Saidpur and other locations, mirroring the national concentration of higher education institutions in the capital. Just over half of the participants (n = 28) reported no formal training in assessment, while 24 indicated some form of assessment-related training or coursework. Given that teachers may acquire assessment skills informally through experience, mentorship, self-study, a lack of formal training does not necessarily imply a lack of ability. However, it does raise questions about how consistently they have been supported to develop a well-rounded assessment literacy that goes beyond designing exams to include feedback and data use with stakeholders. The subsequent sections examine how this mixed preparation profile is reflected in self-reported Skill and Use scores.

Table 1. Participant Characteristics (N = 52)

Characteristic	Category	N
Institution Type	Private	33
	Public	19
Formal Training	No	28

in Assessment	Yes	24
Gender	Man	30
	Woman	21
	Prefer not to say	1
Region	Dhaka	47
	Khulna	2
	Other	3

Note. Data derived from the participant information section of the survey.

Table 2 presents the descriptive statistics for the participants' years of teaching experience. The mean length of experience was 5.5 years. However, the median was considerably lower at 2.25 years, indicating that the experience distribution was positively skewed, with the mean substantially exceeding the median by 3.25 years.

Table 2. Descriptive Statistics for Years of Teaching Experience

N	Mean	Median	SD	Minimum	Q1	Q3	Maximum

52	5.50	2.25	6.05	0.20	1.48	10.00	27.00

Note. Experience is reported in years. Q1 = 1st Quartile; Q3 = 3rd Quartile.

Teaching experience showed marked heterogeneity within the sample, with a mean of 5.5 years but median of only 2.25 years, indicating a strongly positively skewed distribution. As shown in Table 2, half the sample (n = 26) had 1.48 years or less of experience, while the interquartile range spanned 8.52 years (from Q1 = 1.48 to Q3 = 10.00), and the distribution extended from 0.2 years to 27 years. This wide spread means the sample included both relative newcomers learning institutional assessment norms and experienced practitioners who had refined their approaches over decades, creating substantial internal variability. Consequently, the findings are primarily informative about assessment literacy among early-career lecturers in their first few years of university teaching rather than representative of the profession more broadly. Therefore this heterogeneity is both a strength and a limitation. On one hand, the prevalence of teachers with less than two years of experience (n = 26, or 50% of the sample) suggested that the Skill and Use ratings are particularly informative about how early-career lecturers perceive their own competence and what practices they have already begun to enact or avoid within their first years at university. This is analytically valuable in an underresearched context where teacher induction and mentoring in assessment may be limited. On the other hand, the presence of a small number of very experienced teachers means that group-level findings (e.g., overall mean Skill = 3.71) are not representative of a typical lecturer. The outlier with 27 years of experience, for instance, contributed disproportionately to the group mean, even though they may be an exceptional case. When interpreting whether a mean Skill or Use score reflects the "typical" teacher in this sample,

the median and quartile values are often more informative than the mean.

4.2.2 Overall and Subdomain Summaries for Skill and Use

Participants rated their perceived Skill and their frequency of Use for 67 assessment practices using parallel five-point scales (1 = not at all skilled/never, 5 = highly skilled/very often). As shown in Table 3, composite scores for the overall scales and the seven subdomains revealed that teachers reported moderate Skill (M = 3.71, SD = 0.63) and slightly lower Use (M = 3.44, SD = 0.57). Figures 1 and 2 display the distributions of these composite scores, showing that most respondents clustered near the midpoint of the 1–5 range, with relatively limited dispersion around the mean. The violin plot (Figure 3) further illustrated this concentration, with medians positioned slightly below the centre of the scale for both Skill and Use.

This clustering around mid-scale values warranted critical interpretation. While it may reflected genuine moderate competence among early-career teachers (Zhang & Burry-Stock, 2003), it could also indicate response acquiescence or central tendency bias, where participants default to neutral or middle options when uncertain or when questions feel context-dependent (Tourangeau et al., 2000; Krosnick, 1991). Such patterns are common in self-report instruments, particularly when respondents lack formal training (as 54% of this sample did) and may not have the metacognitive frameworks to accurately assess their own competence. This is a phenomenon consistent with the Dunning–Kruger effect, whereby individuals with limited expertise struggle to evaluate their true skill level (Kruger & Dunning, 1999). Additionally, social desirability bias may inflate self-ratings (Paulhus & Vazire, 2007), meaning that what appears as "moderate skill" could mask deeper gaps between perceived and actual competence.

The Skill–Use relationship revealed in Table 3 is more nuanced than a simple deficit model

would suggest. Teachers may possess theoretical knowledge of assessment practices but fail to enact them because of structural and institutional constraints rather than individual incompetence (DeLuca et al., 2016; Xu & Brown, 2016). Time pressures, large classes (often exceeding 100 students), restrictive curriculum mandates, and centralised examination systems can prevent even well-prepared teachers from implementing formative or performance-based assessments (Brookhart, 2011). The qualitative responses in this study repeatedly emphasised such barriers, suggesting that knowing how to conduct an assessment and being institutionally positioned to do so are often decoupled in resource-constrained higher education contexts (Stiggins, 1991; Black & Wiliam, 1998).

Subdomain analysis (Table 3) revealed a stratified profile that reflected the institutional priorities of Bangladeshi universities. The highest means for both Skill (M = 3.89) and Use (M = 3.57) occurred in Construction of Traditional Tests, while Ethics and Integrity showed similarly elevated Skill (M = 3.88) and Use (M = 3.69) ratings. These domains align with an exam-centric culture where teachers are routinely responsible for writing, administering, and marking high-stakes tests while managing confidentiality (Stiggins, 1995). This pattern is unsurprising given that summative, written examinations dominate assessment in Bangladeshi higher education, often at the expense of formative feedback and alternative assessment modes (Rahman & Pandian, 2018).

Conversely, the lowest means appeared for Standardised Testing and Data Use (Skill M = 3.53, Use M = 3.28) and Communication of Results (Skill M = 3.54, Use M = 3.17), domains requiring technical statistical literacy and sustained engagement with diverse stakeholders (parents, administrators, external bodies). These findings mirror international research showing that teachers often lack training in psychometric interpretation and data-driven decision-making

(Popham, 2009; Brookhart, 2011). Critically, these low-scoring domains were precisely where teachers in open-ended responses requested further training, mentioning item analysis, data-informed feedback, and systematic communication protocols. This convergence between quantitative gaps and qualitative needs suggests that the absence of formal preparation in these technical areas is perceived by teachers themselves as a barrier to effective practice (Zhang & Burry-Stock, 2003).

The overall descriptive statistics suggested that the teachers exhibited relative strength in traditional testing and rule compliance and skills reinforced through routine practice and showed weaker self-reported competence in data interpretation and stakeholder communication. The weaker areas are rarely emphasised in pre-service or in-service professional development (DeLuca et al., 2016; Xu & Brown, 2016). The following sections examine the reliability of these scales and the magnitude of Skill–Use gaps before turning to group differences and qualitative themes that illuminate these quantitative patterns.

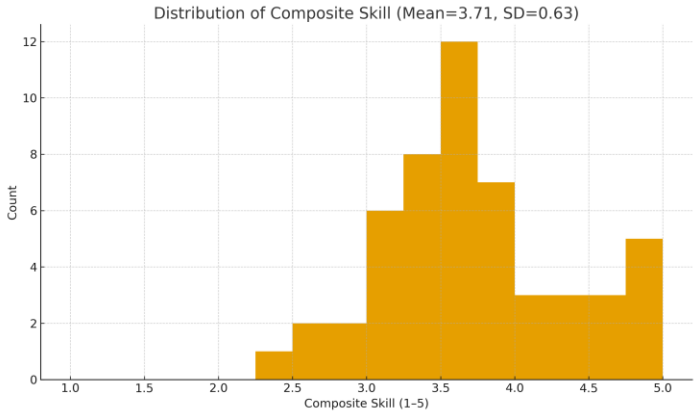


Figure 1: Histogram of composite Skill scores (1–5) computed as the participant-level mean across 67 Skill items.

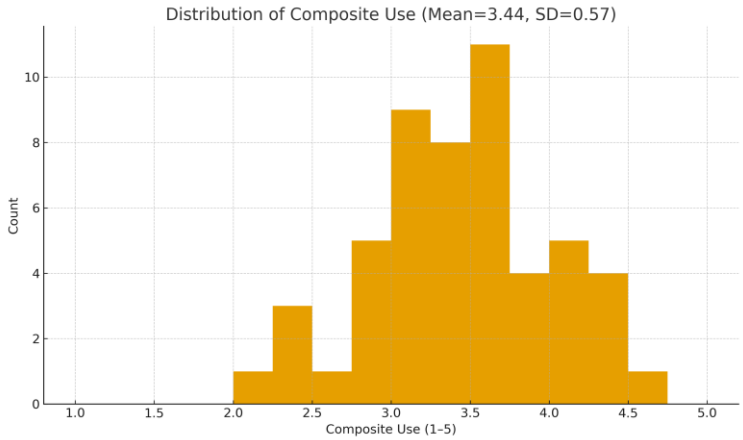


Figure 2: Histogram of composite Use scores (1–5) computed as the participant-level mean across 67 Use items.

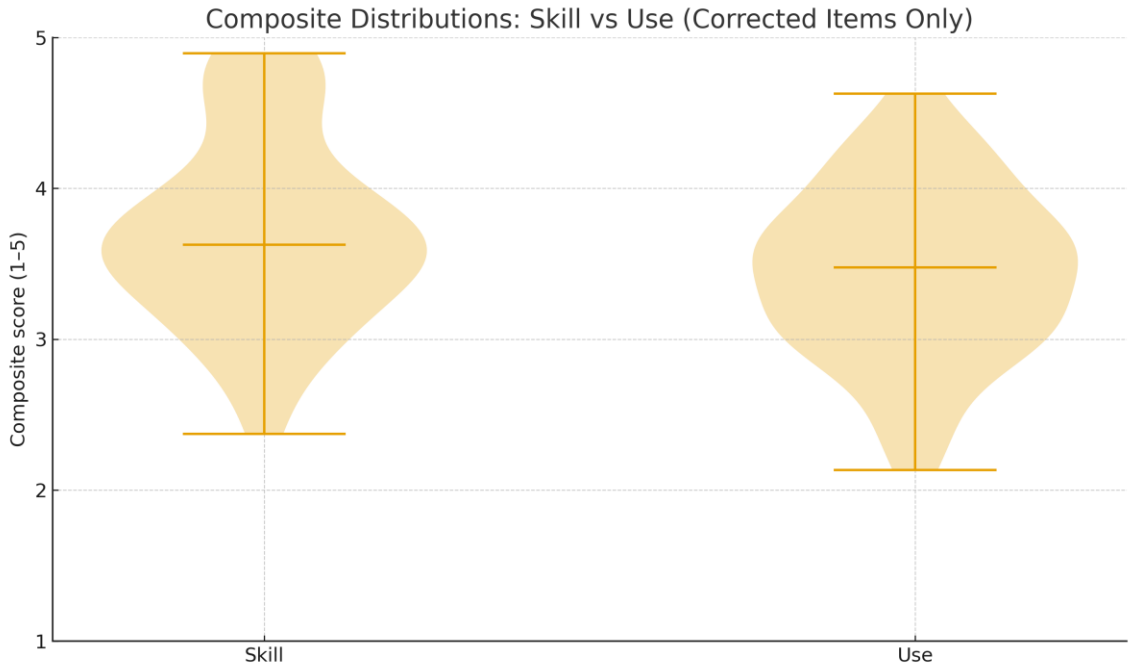


Figure 3: Violin plots for Skill and Use composites (1–5). Y-axis constrained to the response range; medians shown as horizontal lines.

Table 3. Descriptive Statistics for Overall and Subdomain Scores for Skill and Use

Domain/Subdomain	Skill (M)	Skill (Mdn)	Skill (SD)	Use (M)	Use (Mdn)	Use (SD)
Overall (67 items)	3.71	3.63	0.63	3.44	3.48	0.57
Selection & Planning	3.78	3.78	0.57	3.63	3.56	0.52
Construction of Traditional Tests	3.89	3.8	0.69	3.57	3.6	0.53
Performance-Based Assessment	3.72	3.74	0.69	3.43	3.56	0.72
Standardised Testing & Data Use	3.53	3.58	0.84	3.28	3.33	0.85
Grading & Feedback Practices	3.71	3.45	0.74	3.44	3.36	0.68
Communication of Results	3.54	3.6	0.81	3.17	3	0.65
Ethics & Integrity	3.88	4	0.83	3.69	3.67	0.99

Note. Scores are mean ratings on a 1–5 scale. *M* = Mean; *Mdn* = Median; *SD* = Standard Deviation.

4.3 Reliability of the Assessment Literacy Scales

Internal consistency reliability was assessed using Cronbach's alpha (α) for the overall Skill and Use scales and for each of the seven subdomains. Cronbach's alpha estimates the proportion of variance in observed scores attributable to true score variance rather than random error, with values of $\alpha \geq .70$ typically regarded as acceptable and values above .90 often described as excellent in conventional guidelines (Nunnally & Bernstein, 1994; George & Mallery, 2016).

Table 4 reports alpha coefficients for all scales. All Skill subscales showed acceptable to excellent internal consistency ($\alpha = .71–.92$), and five of the seven Use subscales met the .70 threshold. Two Use subscales, Construction of Traditional Tests ($\alpha = .64$) and Communication of Results ($\alpha = .54$) fell below this threshold, indicating that items within these domains were answered less consistently and may be capturing a broader or more heterogeneous set of practices. These subscales should therefore be interpreted cautiously, as their composite scores may be less stable.

The overall Skill scale (67 items) yielded $\alpha = .97$ and the overall Use scale $\alpha = .96$. On the surface, these values exceed the usual benchmark for excellent reliability. However, values approaching 1.0 require careful interpretation. Reliability coefficients of this magnitude may reflect methodological artefacts. Three considerations are important.

First, data-quality checks using the cleaned dataset showed that 29 of 52 participants (55.8%) had a standard deviation below 0.5 across the 67 Skill items, and 27 of 52 (51.9%) had $SD < 0.5$ for the 67 Use items. Such low within-person variability indicated that many teachers gave similar ratings to a large number of items (for example, repeatedly selecting "3 = moderately skilled" or "4 = often"). It is a pattern of straightlining or response sets rather than fine-grained differentiation between practices (DeSimone et al., 2015).

Second, the mean inter-item correlations were approximately $r \approx .36$ for the Skill scale and $r \approx .25$ for the Use scale. Inter-item correlations of this magnitude, when combined with a large number of items, indicate that many items are tapping closely related behaviours and beliefs, raising the possibility of content overlap and redundancy (Cortina, 1993; Streiner, 2003). Under such conditions, Cronbach's alpha is mathematically inflated by both high inter-item correlations and the length of the scale (Cortina, 1993; Streiner, 2003).

Third, the discrepancy between the very high overall alphas and the more modest (and sometimes low) subscale alphas indicates that the instrument is not a simple unidimensional scale. The variation in subscale reliability ($\alpha = .54–.92$) provided some evidence that different facets of assessment literacy are being captured, but the overall coefficients are likely to reflect a mixture of true consistency and response similarity across items rather than "perfection" in measurement (Tavakol & Dennick, 2011).

In light of these observations, the overall alphas ($\alpha = .97$ and .96) should be interpreted with caution. They confirm that responses are highly consistent at a formal statistical level, but they also suggest that this consistency is partly driven by limited variability in how participants used the response scales and by redundancy across items. The cleaned file excluded incomplete and obviously inconsistent cases, yet the remaining patterns indicated that some teachers responded in broadly similar ways across many practices, which may be a realistic reflection of moderate, generalised confidence rather than detailed discrimination between specific assessment competencies. A detailed discussion of the limitations of the very high alpha values is in Section 5.2. Despite these measurement caveats, the overall scales remained suitable for addressing the research questions, as the focus was on identifying patterns and gaps across the sample rather than on precise individual-level measurement of assessment competence.

Table 4. Internal Consistency of Assessment Literacy Scales (Cronbach's Alpha)

Scale/Subscale	Number of Items (*k*)	Cronbach's Alpha (α)
Overall Skill	67	.97
Overall Use	67	.96
Skill Subscales		
Selection & Planning	9	.81
Construction of Traditional Tests	10	.86
Performance-Based Assessment	17	.92
Standardised Testing & Data Use	12	.92
Grading & Feedback Practices	11	.90

Communication of Results	5	.78
Ethics & Integrity	3	.71
Use Subscales		
Selection & Planning	9	.74
Construction of Traditional Tests	10	.64
Performance-Based Assessment	17	.91
Standardised Testing & Data Use	12	.90
Grading & Feedback Practices	11	.83
Communication of Results	5	.54
Ethics & Integrity	3	.75

4.4 Scale level Wilcoxon signed-rank test with Benjamini–Hochberg false discovery rate (FDR) correction

4.4.1 Methodological Framework

This analysis compared teachers' self-reported skill ("Skill") with their reported use ("Use") of assessment practices at the scale level (seven API domains and the overall composite). The Wilcoxon signed-rank test served as the non-parametric analogue of the paired-samples t-test, appropriate for ordinal Likert-type responses and when the distribution of paired differences cannot be assumed normal (Cohen et al., 2018; Conover, 1999; Field, 2018; Wilcoxon, 1945). The test evaluated whether the within-teacher median difference (Skill − Use) exceeded zero. A directional alternative (Skill > Use) was specified a priori, motivated by prior evidence that teachers often report stronger assessment knowledge than is executed in everyday practice (Stiggins, 1991; Brookhart, 2011; DeLuca et al., 2016).

Because seven domain tests were conducted, the probability of false positives was managed using the Benjamini–Hochberg false discovery rate (FDR) procedure at q = .05 (Benjamini & Hochberg, 1995; Benjamini, 2010). Adjusted p-values below .05 were treated as statistically significant. Pairs with a zero Skill–Use difference within a domain were omitted automatically by the Wilcoxon algorithm; the results tables report the number of non-zero pairs analysed for each test.

Effect sizes were expressed as $r = Z/\sqrt{N}$, where Z is the standardised Wilcoxon statistic and N is the number of non-zero paired observations analysed for that domain. Two-tailed tests were used. Unless stated otherwise, all p-values are Benjamini–Hochberg FDR-adjusted at q = .05. The resulting r can be interpreted like a correlation (rank biserial), with approximate thresholds of .10 (small), .30 (medium), and .50 (large) used to aid practical interpretation (Cohen, 1988;

Tomczak & Tomczak, 2014). Alongside significance and effect size, the tables report the mean Skill–Use gap for each scale to support a substantive reading of the results. For transparency, secondary item-level checks were carried out only within domains that were significant after FDR control, with within-domain FDR at q = .05; these summaries appear later, and full details are provided in Table 10 of Appendix 2.

4.4.2 Results by scale

Table 1 summarises the findings. All p values reported in this section were adjusted using the Benjamini–Hochberg FDR at q = .05. The overall composite score showed higher Skill than Use (Mdiff = +0.27), Wilcoxon W = 103.5, $p < .001$, rank biserial r = .81. Across the seven domains, every scale showed a significant Skill > Use gap after FDR control (all $p \leq .008$). The largest differences were in Communication of Results (Mdiff = +0.38, r = .95), Construction of Traditional Tests (Mdiff = +0.33, r = .78), Performance-Based Assessment (Mdiff = +0.29, r = .77), and Grading and Feedback Practices (Mdiff = +0.27, r = .89). Selection and Planning showed a smaller but reliable gap (Mdiff = +0.15, $p < .001$).

From a measurement perspective, the pattern was coherent with earlier internal consistency evidence for the composites and with arguments that aggregated scores provide more stable estimates of underlying practice and knowledge (Nunnally & Bernstein, 1994). From a practical perspective, the largest gaps clustered in domains that required time, collaboration, feedback, and dialogue (for example, feedback cycles and communicating results). This aligned with the literature that described a recurrent distance between assessment knowledge and routine enactment in classrooms (Black & Wiliam, 1998; Brookhart, 2011; DeLuca et al., 2016; Stiggins, 1991). The results suggested that respondents felt confident about what good practice

looked like yet faced barriers to using it consistently. In contrast, Selection and Planning showed a smaller gap, which suggested that planning knowledge translated more readily into daily work.

A critical reading considered both statistical and practical significance. Rank biserial values around .50 are usually interpreted as moderate, and values above .70 as large (Tomczak & Tomczak, 2014). The effects reported here were generally large. However, the Wilcoxon test reduced the analytic sample within each domain by removing zero-difference pairs, which may have inflated effect sizes slightly in smaller domains. The scale differences were also self-reported and therefore may reflect optimism about competence or under-reporting of frequency when time or institutional constraints were salient. Despite these caveats, the consistent direction and the corrected significance across all domains strengthened the inference that the Skill–Use gap was a broad feature rather than an artefact of one or two domains.

Table 5. Wilcoxon signed-rank tests comparing Skill vs Use by domain (FDR-adjusted)

Domain	**n (non-zero pairs analysed)**	**Skill M**	**Use M**	**Mean diff**	**W**	**p (unadjusted)**	**p (FDR-adjusted)**	**r (rank biserial)**
Overall composite (67 items) *	46	3.71	3.44	0.27	103.5	< .001	—	.81
Communication of Results	27	3.54	3.17	0.38	10.0	< .001	< .001	.95

Construction of Traditional Tests	35	3.89	3.57	0.33	69.0	< .001	< .001	.78
Performance-Based Assessment	41	3.72	3.43	0.29	99.0	< .001	< .001	.77
Grading and Feedback Practices	32	3.71	3.44	0.27	29.5	< .001	< .001	.89
Standardised Testing and Data Use	37	3.53	3.28	0.25	91.0	< .001	.001	.74
Ethics and Integrity	19	3.88	3.69	0.19	29.0	.008	.008	.69
Selection and Planning	34	3.78	3.63	0.15	101.5	< .001	< .001	.66

Note. n (pairs analysed) = number of non-zero pairs used by the Wilcoxon test; N = 52 respondents. M = Mean. W = Wilcoxon statistic (two-tailed). p (unadjusted) = p before adjustment. p (FDR-adjusted) = Benjamini–Hochberg adjusted p across the seven domain tests at q = .05. r = rank-biserial correlation. *The overall test is descriptive and was not included in the FDR family for domains.

The largest gaps appeared in domains requiring time and institutional support (Communication, Performance-Based Assessment), while Selection & Planning showed smallest gap (Mdiff=0.15), suggesting planning knowledge translates more readily to practice.

4.5 Differences by Institution Type, Prior Assessment Training and Teaching Experience

4.5.1 Institution Type Comparisons

The Mann–Whitney U tests revealed no statistically significant differences in assessment literacy scores between teachers from public and private universities across any of the seven domains (see Table 6). All p-values were well above the conventional 0.05 threshold (ranging from .226 to .962) (Cohen et al., 2018). The corresponding effect sizes were uniformly small, consistent with accepted benchmarks for r and Cliff's δ (Cohen, 1988; Romano et al., 2006). For example, the largest observed difference was in the Selection & Planning domain, where public university teachers had slightly higher median scores than their private university counterparts (U = 249.5, p = .226, Cliff's δ = –0.20). But this difference was not statistically significant and represented only a small effect. Similarly, a small non-significant advantage for public institution teachers was noted in Performance-based Assessment (U = 254.5, p = .266, δ = –0.19). Overall, the Cliff's δ values ranged from –0.01 to –0.20 (absolute values), indicating negligible practical differences between the two institution types. This suggests that institutional context (public vs. private) was not associated with any meaningful variation in the assessment literacy of the teachers in this sample.

Table 6. Mann–Whitney U Test Results: Institution Type Comparison

Assessment Domain	Measure	U	p	Cliff's δ	Effect Size
Communication & Ethics	Skill	283.0	.567	–.097	Small
Communication & Ethics	Use	257.0	.286	–.18	Small
Grading Practices	Skill	281.5	.549	–.102	Small
Grading Practices	Use	303.5	.857	–.032	Small
Interpretation of Standardised Tests	Skill	303.5	.856	–.032	Small
Interpretation of Standardised Tests	Use	285.0	.593	–.091	Small
Performance-based Assessment	Skill	254.5	.266	–.188	Small
Performance-based Assessment	Use	263.5	.345	–.159	Small
Selection & Planning	Skill	249.5	.226	–.204	Small
Selection & Planning	Use	297.5	.767	–.051	Small
Traditional Test Construction	Skill	286.0	.607	–.088	Small
Traditional Test Construction	Use	286.5	.614	–.086	Small
Use of Assessment Results	Skill	296.0	.746	–.056	Small
Use of Assessment Results	Use	310.5	.962	–.01	Small

Note. Private institutions: $n = 33$; Public institutions: $n = 19$. Negative Cliff's δ values indicate higher scores for public institution teachers. No comparisons were statistically significant at the 0.05 level.

4.5.2 Formal Training Comparisons

Comparisons based on prior assessment training (formal training *vs.* no formal training) showed a consistent pattern of higher scores among the teachers who had received formal training. Although these differences did not reach statistical significance after correcting for multiple comparisons. As Table 7 implies, two domain-level differences were nominally significant before adjustment: teachers with formal assessment training scored higher in *Grading Practices – Skills* ($U = 228.5$, $p = .049$, Cliff's $\delta = –0.32$) and in *Use of Assessment Results – Use* ($U = 223.0$, $p = .038$, $\delta = –0.34$), reflecting medium effect sizes in favour of the trained group. However, after controlling the false discovery rate, neither of these differences remained statistically significant (adjusted $q > .05$). In addition, several other domains exhibited positive trends ($p < .10$) associated with training. Notably, the *Communication & Ethics – Skills* domain showed a marginal difference ($p = .059$, $\delta = –0.31$), as did *Selection & Planning* (for both skills and use, $p \approx .07–.08$, $\delta \approx –0.29$) and *Use of Assessment Results – Skills* ($p = .081$, $\delta = –0.28$). Although these did not reach the $p < .05$ criterion, all of the corresponding effect sizes were in the small-to-medium range and consistently negative, indicating higher median scores for teachers with formal training across all domains. In fact, every Cliff's δ was negative in Table 9, suggesting a uniform trend whereby trained teachers outperformed untrained teachers on assessment literacy measures, even if the differences were generally modest. This trend exerts the practical importance of formal assessment training: teachers who had received structured training tended to report greater assessment-related skills and more frequent use of sound assessment practices (especially in areas like performance-based assessment, use of results, and grading), compared to those without training. It should be noted that the variability was considerable, and the differences did not meet the threshold for statistical significance once the number of comparisons was taken into account.

Table 7. Mann–Whitney U Test Results: Formal Training Comparison

Assessment Domain	Measure	U	p	Cliff's δ	Effect Size	Sig.
Communication & Ethics	Skill	233.0	.059	–.307	Medium	†
Communication & Ethics	Use	265.5	.197	–.21	Small	ns
Grading Practices	Skill	228.5	.049	–.32	Medium	*
Grading Practices	Use	251.5	.123	–.251	Small	ns
Interpretation of Standardised Tests	Skill	250.5	.117	–.254	Small	ns
Interpretation of Standardised Tests	Use	278.5	.293	–.171	Small	ns
Performance-based Assessment	Skill	269.5	.225	–.198	Small	ns
Performance-based Assessment	Use	300.0	.513	–.107	Small	ns
Selection & Planning	Skill	240.0	.079	–.286	Small	†
Selection & Planning	Use	239.0	.075	–.289	Small	†
Traditional Test Construction	Skill	298.5	.496	–.112	Small	ns
Traditional Test Construction	Use	315.0	.706	–.062	Small	ns
Use of Assessment Results	Skill	240.5	.081	–.284	Small	†
Use of Assessment Results	Use	223.0	.038	–.336	Medium	*

Note. No formal training: $n = 28$; Formal training: $n = 24$. *$p < .05$, † $p < .10$ (two-tailed); ns = not significant. Negative Cliff's δ values indicate higher scores for teachers with formal training.

In summary, while none of the training-related differences was statistically confirmed at the 5% level after adjustment, the pattern of results suggests that formal assessment training may confer

practical benefits. The trained group tended to score higher, particularly in fundamental domains such as grading practices and using assessment results to inform teaching. These findings are consistent with the idea that structured professional development can enhance teachers' assessment competencies in meaningful ways, even if the magnitude of improvement varies by domain. The lack of a uniform significant effect across all areas also implies that certain aspects of assessment literacy (for instance, highly specialised skills like standardised test interpretation) might not be fully addressed by general training programmes and could require additional support or targeted training (Brookhart, 2011; DeLuca et al., 2016).

4.5.3 Teaching Experience Correlations

Spearman's rho analyses revealed significant positive correlations between years of teaching experience and teachers' assessment literacy scores on several domains (Table 10). In particular, more experienced teachers reported substantially higher self-perceived skills in *Grading Practices* ($\rho = .482$, $p < .001$) and *Performance-based Assessment* ($\rho = .467$, $p = .001$). These correlations represent large effect sizes. The implication is that competence in designing fair grading systems and implementing performance-based assessments tends to improve markedly with greater teaching experience. Significant medium-strength correlations were also found for a number of other domains. For example, experience was positively associated with *Use of Assessment Results – Skills* ($\rho = .405$, $p = .003$) and *Performance-based Assessment – Use* ($\rho = .403$, $p = .003$), along with overall assessment literacy ($\rho = .447$ for overall skill, $\rho = .353$ for overall use, both $p \leq .01$). This suggests that veteran teachers feel more skilled in assessment and tend to make somewhat more frequent use of effective assessment practices in their classrooms. In contrast, a few domains showed little to no relationship with experience. Notably, there were no significant correlations for *Traditional Test Construction* ($\rho \approx .21$, $p > .13$) or *Interpretation*

of Standardised Tests ($\rho \approx .22$, $p > .11$), implying that simply spending more years in the profession does not automatically improve teachers' abilities in those more technical or externally oriented assessment areas. It is also worth mentioning that the correlations with teachers' "frequency of use" of assessment strategies were generally weaker than those for "skill" perceptions. For instance, years of experience had a more substantial effect on teachers' self-rated *Grading* and *Assessment design skills* than on how often they reported using those assessments in practice (compare $\rho = .482$ vs $.365$ for Grading Practices skill vs. use). This outlines that experience contributes more to teachers' confidence and knowledge in assessment (their perceived skill) than to changes in their day-to-day assessment behaviour. In fact, only one correlation involving usage frequency failed to reach significance at the 5% level. *Communication & Ethics – Use* ($\rho = .254$, $p = .072$) was a marginal case, whereas the corresponding skill for that domain was significant ($\rho = .278$, $p = .048$). This disparity between skill and use may indicate that teachers feel more adept at various assessment tasks. But there are certain constraints or habits which do not always let them translate that greater competence into action in the classroom.

Table 8. Spearman's ρ Correlations: Assessment Literacy and Teaching Experience

Assessment Domain	Measure	Spearman's ρ	p	Effect Size	Sig.
Communication & Ethics	Skill	.278	.048	Small	*
Communication & Ethics	Use	.254	.072	Small	†
Grading Practices	Skill	.482	< .001	Large	***
Grading Practices	Use	.365	.008	Medium	**

Assessment Domain	Measure	Spearman's ρ	p	Effect Size	Sig.
Interpretation of Standardised Tests	Skill	.216	.128	Small	ns
Interpretation of Standardised Tests	Use	.221	.118	Small	ns
Performance-based Assessment	Skill	.467	.001	Large	**
Performance-based Assessment	Use	.403	.003	Medium	**
Selection & Planning	Skill	.357	.010	Medium	*
Selection & Planning	Use	.348	.012	Medium	*
Traditional Test Construction	Skill	.215	.131	Small	ns
Traditional Test Construction	Use	.107	.454	Small	ns
Use of Assessment Results	Skill	.405	.003	Medium	**
Use of Assessment Results	Use	.330	.018	Small	*
Overall Assessment Literacy	Skill	.447	.001	Large	**
Overall Assessment Literacy	Use	.353	.011	Medium	*

Note. $N = 51$. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .10$ (two-tailed).

4.6 Qualitative Analysis of Open-Ended Responses

4.6.1 Overview

The two open ended questions invited participants to describe challenges in assessing students effectively and to suggest types of assessment related training that would benefit them most. Forty two participants responded to the question on challenges and forty one to the question on

training needs. The responses were typically one to three sentences long, providing concise but focused insights into how teachers experienced assessment in their institutional contexts.

Five overarching themes were identified:

1. Contextual constraints (time, class size, curriculum and institutional policies)
2. Student related factors (preparedness, engagement, and academic integrity)
3. Lack of assessment specific training and institutional support
4. Tension between traditional examinations and alternative assessments
5. Technical and digital skills gaps in using assessment data and tools

Table 9 summarises these themes and subthemes. The sections below describe each theme, provide aligned quotations, and indicate where themes help interpret the quantitative findings on the Skill–Use gap and subdomain patterns.

Table 9. Themes and Subthemes Coded from Content Analysis

Theme	Subthemes	Illustrative Focus
Contextual constraints: time, class size, policy	Large classes and heavy marking; Short semesters and extensive syllabuses; Restrictive assessment outlines and institutional rules	Limited time and high workload make it difficult to implement time intensive assessment and feedback practices.

Student factors: preparedness, engagement, integrity	Variable prior preparation; Limited study effort and exam anxiety; Cheating and unfair means (including AI use)	Differences in student readiness and behaviour complicate fair assessment and discourage more open–ended tasks.
Lack of assessment training and institutional support	Absence of formal assessment training; Learning assessment “by oneself”; Limited institutional guidance or resources	Teachers feel expectations for assessment are not matched by systematic professional development or support.
Traditional examinations vs alternative assessments	Dominance of summative written exams; Exam oriented culture; Interest in formative and performance based approaches	Lecturers recognise the value of more diverse assessment methods but feel constrained by exam focused norms and policies.
Technical and digital skills gaps	Limited skills in data analysis and interpretation; Challenges with digital platforms and plagiarism detection tools	Teachers would like to use assessment data and technology more systematically but feel underprepared to do so.

4.6.2 Contextual Constraints (Time, Class Size, Policy)

Many participants described structural conditions that made it difficult to implement assessment practices beyond basic examination formats. A recurring concern was large class size combined

with short semesters and heavy marking loads. One teacher summarised this by stating that “a huge number of exam scripts to check for each course” (Participant 9) leaves little room for more detailed assessment or feedback. Another explained that “extensive syllabus leading to time constraints and tendencies of academic dishonesty among students” (Participant 34) meant that even when they wanted to assess more thoughtfully, they could not do so consistently.

Institutional policies and expectations also appeared as constraints. Several participants referred to restrictive assessment outlines or expectations that limit the range of permissible methods, such as pressure to adhere closely to predetermined exam formats or to maintain high pass rates. One lecturer noted that “class size, administrative responsibilities, frequent changes in institutional rules, extremely tight deadlines, [and] restrictive assessment expectations” (Participant 27) collectively narrowed the space for experimentation with alternative assessment. These contextual constraints help explain why the largest Skill–Use gaps in the quantitative results appeared in domains that are time intensive and interactional, such as performance based assessment, grading and feedback practices, and communication of results: teachers may feel reasonably skilled but lack realistic opportunities to enact these practices fully in large, exam oriented classes.

4.6.3 Student Factors (Preparedness, Engagement, Academic Integrity)

A second theme concerned student related factors that teachers felt affected the fairness and effectiveness of assessment. Several participants highlighted variability in students’ prior preparation. One explained that “the students do not have the basic understanding of the subject they are studying. Hence, it gets difficult to provide them with higher levels of knowledge which is norm in university level. A standard assessment will lead to student failing in class”

(Participant 4). Another pointed to mixed ability groups and the challenge of setting standards that are fair to both weaker and stronger students (Participant 7).

Low engagement and exam anxiety were also mentioned. Some teachers described students who "do not put effort in studying at all and cannot accept what assessment they got for not learning or studying" (Participant 25). Others noted that test anxiety or personal circumstances could depress performance on high stakes tasks. Academic dishonesty emerged as a prominent concern, particularly in large classes and online or take home contexts. One respondent pointed to "new techniques of unfair means in exam" (Participant 18) and the use of artificial intelligence to prepare assignments, while another emphasised difficulty ensuring that grammatical and reading tasks were completed honestly. These student related factors resonate with the quantitative finding that teachers reported relatively lower use of performance based and alternative assessments: when cheating, rote learning, or uneven preparation are perceived as prevalent, teachers may feel compelled to rely on tightly controlled written exams even if they recognise the value of more authentic tasks.

4.6.4 Lack of Assessment Training and Institutional Support

A third theme centred on limited formal preparation in assessment and the perceived need for more structured professional learning. Several participants explicitly stated that they had "no formal training in assessment" (Participant 13) and had largely "learnt by myself" (Participant 39). Others reported that while they had some general pedagogical training, assessment had been only briefly covered. One respondent reflected that "teachers responsibility should not be solely assumed without adequate training and institutional support," emphasising that expectations for sophisticated assessment practice were not matched by opportunities to develop the necessary expertise.

Requests for training were concrete rather than abstract. Participants expressed interest in sessions on grading and evaluation, constructing fair and valid exam papers, using data to diagnose learning needs, and giving effective feedback. For example, one teacher mentioned wanting training on "how to use data based informatics to get a better understanding of the quality of education," while another asked for "professional training" and "university wide training on assessment" (Participants 9, 12, 23, 44, 48). These comments align with the quantitative pattern that teachers with formal training tended to report somewhat higher scores, particularly in domains like grading practices and use of assessment results, even when differences did not always reach statistical significance. They also support the interpretation that self-reported Skill scores may partly reflect confidence built through practice and informal learning, whereas targeted training could help deepen understanding in more technical areas such as data use and communication of results. Assessment knowledge remains "generally inadequate relative to standards and expectations", and targeted professional development is needed (Xu & Brown, 2016, p.14). Participants' voices strongly support this need.

4.6.5 Traditional Examinations and Alternative Assessments

Many participants described a tension between the dominant culture of summative examinations and their interest in more varied assessment approaches. Several respondents characterised their institutional environment as "exam oriented," noting that formal examinations continue to "dominate" assessment and that syllabuses and assessment policies often prioritise coverage and grading over formative feedback. One lecturer commented that the "challenges of the syllabuses and its needs for the students coming from different backgrounds cause a bit trouble while assessing," (Participant 25) indicating that prescribed exam formats do not always match students' learning needs.

At the same time, teachers expressed a desire for training and support in alternative assessment methods. Suggested topics included “performance assessment tools,” “portfolio based assessment,” “peer assessment,” and “assessment integrated learning.” One participant wanted help with “tech enhanced assessment tools,” while another asked for guidance on “student centred, engaging and inclusive assessments.” These aspirations provide qualitative support for the quantitative finding that skill ratings for performance based assessment and grading and feedback practices were moderate, but use ratings were lower. Teachers appear to recognise the value of more formative, student centred assessment and feel moderately capable conceptually, yet structural and cultural pressures keep practice anchored in traditional written exams.

This tension between traditional examination culture and interest in alternative assessment is widely documented in higher education contexts where summative exams dominate institutional assessment policies (Brookhart, 2013; Carless, 2015). In exam-oriented systems, teachers often feel caught between their pedagogical knowledge of what constitutes good practice and the institutional expectations or accreditation requirements that mandate specific assessment formats (Alam et al., 2020). The qualitative responses in this study reflect this tension and provide a plausible explanation for why Construction of Traditional Tests showed relatively strong Skill and Use alignment (both scores above 3.5), whereas Performance-Based Assessment and Communication of Results showed larger gaps: traditional exams are institutionally mandated and routinized, whereas alternative methods require discretionary time, support, and cultural acceptance that may not be available (Black & Wiliam, 1998; DeLuca et al., 2016).

4.6.6 Technical and Digital Skills Gaps

The final theme involved perceived gaps in technical and digital skills related to assessment. Some participants mentioned difficulties in analysing assessment data systematically, such as conducting item analysis or interpreting test statistics. One noted a wish for training in “statistical analysis of test” (Participant 1) results, and another linked their training needs to “data based informatics” for understanding teaching quality. Others referred to challenges in using plagiarism detection tools or managing technology enhanced assessment platforms (Participant 32, 38).

These comments correspond with the relatively low means and sizeable Skill–Use gaps in the “standardised testing and data use” and “communication of results” domains. While not all participants are expected to conduct complex psychometric analyses, several indicated that they would like to move beyond intuitive judgments toward more systematic use of assessment evidence. The fact that these domains also had some of the lowest Use scores suggests that even when teachers feel they understand the basic ideas, they may lack the software skills, time, or institutional infrastructure to apply them regularly in their own courses.

Limited technical skills in assessment data use are a common finding in teacher assessment literacy research, particularly in contexts where preservice training emphasizes classroom management and lesson design over psychometric principles or data-driven decision-making (Popham, 2009; Brookhart, 2011). The fact that Standardised Testing and Data Use had the lowest mean Skill ($M = 3.53$) and Use ($M = 3.28$) scores in the quantitative analysis (Table 3), combined with explicit requests for training in "statistical analysis" and "data-based informatics" in the qualitative responses, suggests that this is an area where many participants lack both conceptual knowledge and practical tools. This pattern reinforces the interpretation that some

domains of assessment literacy are less developed than others, not only because of limited training but also because institutional infrastructure (e.g., item analysis software, learning analytics platforms) may not support routine data use even when teachers are willing to engage with it (DeLuca et al., 2016).

4.6.7 Reconciling Quantitative Self-Reports with Qualitative Accounts

The qualitative findings provide important context for interpreting the quantitative self-report data. On the surface, there appears to be a tension between teachers' moderate-to-high self-rated Skill scores (overall M = 3.71, with several domains exceeding M = 3.8) and the extensive challenges and training gaps they described in open-ended responses. However, this apparent contradiction can be reconciled when the nature of self-report data and the contextual realities of assessment practice are considered together.

First, self-reported skill ratings reflect teachers' perceived competence based on their awareness of assessment practices, not validated measures of actual proficiency (Kruger & Dunning, 1999; Paulhus & Vazire, 2007). Teachers who lack formal training, as 54% of this sample did, may rate themselves as "moderately skilled" because they have developed workable routines through experience and have not encountered systematic frameworks that would reveal gaps in their knowledge (DeLuca et al., 2016). The qualitative responses support this interpretation: many participants explicitly stated they had "learnt by myself" or had "no formal training in assessment," yet they are expected to design, administer, and grade high-stakes examinations. Their self-ratings may therefore reflect confidence born of necessity and routine practice rather than deep, formally supported competence (Zhang & Burry-Stock, 2003).

Second, the qualitative themes reveal that teachers distinguish between knowing what good assessment looks like and being able to enact it under institutional constraints. Several respondents expressed familiarity with concepts like rubrics, performance assessment, and formative feedback, yet described structural barriers such as large classes, heavy marking loads, restrictive policies that prevented regular implementation. This pattern is consistent with situated models of assessment literacy, which argue that competence is not solely an individual trait but is shaped by the resources, expectations, and cultures of specific institutional contexts (Xu & Brown, 2016). Teachers may feel moderately skilled in the abstract while simultaneously recognizing that their practice is constrained, which helps explain why Use ratings ($M = 3.44$) were consistently lower than Skill ratings.

Third, self-report data are subject to recall bias and social desirability (Schaeffer & Presser, 2003). Teachers may overestimate the frequency with which they provide detailed feedback or use assessment data systematically, particularly if they believe these practices are professionally valued (Paulhus & Vazire, 2007). The qualitative responses, by allowing more nuanced self-reflection, reveal challenges and limitations that teachers might not acknowledge when completing structured rating scales. For example, while a teacher might rate themselves as "skilled" in communicating results to stakeholders, their open-ended comment that "clear expectations and channels" are lacking suggests that their actual practice is more limited than the numeric rating implies.

Chapter 5: Discussion

This study examined the assessment literacy of undergraduate English language teachers in Bangladesh through a cross-sectional survey that combined quantitative and qualitative approaches. Fifty-two teachers from public and private universities completed an adapted Assessment Practices Inventory (API) measuring both self-perceived skill and self-reported frequency of use across seven assessment domains, alongside open-ended questions about challenges and professional development needs. This chapter interprets the key findings in relation to the literature reviewed in Chapter 2, addresses the self-report limitations inherent in the methodology, and explicates the theoretical significance of observed patterns within the Bangladeshi higher education context.

5.1 Theoretical Interpretation of Skill-Use Patterns

5.1.1 The Skill-Use Gap as Situated Compromise Rather Than Knowledge Deficit

The quantitative findings revealed a consistent pattern where teachers reported higher self-perceived Skill (M = 3.71) than self-reported frequency of Use (M = 3.44) across all seven assessment domains. With statistically significant differences after false discovery rate correction (all $p \leq .008$) with medium-to-large effect sizes (r ranging from .44 to .88), the discrepancy is practically meaningful despite the relatively small sample size (N=52). This magnitude indicates that the divergence between Skill-Use is a substantial and practical disconnect which is further discussed in this section. It is important to acknowledge that high Skill ratings may be partially influenced by social desirability bias, as teachers may overestimate their competence to appear professional (Paulhus & Vazire, 2007). However, the fact that these same teachers reported significantly lower Use scores suggests a level of candour; they were willing to admit that they do not frequently practice these methods despite feeling capable. This divergence reinforces the

interpretation that the gap is real and structural, rather than purely an artifact of response bias. This statistical significance could be superficially interpreted as evidence that teachers "know but don't do", meaning they possess assessment knowledge that remains dormant in practice. However, such an interpretation would fundamentally misunderstand both the nature of self-report data and the situated realities of assessment practice in Bangladeshi higher education.

Xu and Brown's (2016) Teacher Assessment Literacy in Practice (TALiP) framework provides a more theoretically sound lens for understanding these patterns. TALiP reconceptualises assessment literacy not as a static knowledge base but as an iterative system in which teachers continuously negotiate between their assessment conceptions and the institutional, policy, and contextual constraints they face. Xu and Brown (2016, p. 157) emphasised that "teachers make compromises when balancing their assessment conceptions against institutional constraints, policy mandates, and stakeholder expectations, resulting in assessment practices that reflect negotiated settlements rather than ideal enactments." This perspective does not reframe the Skill-Use gap observed in this study as individual deficit or implementation failure. Rather it is an evidence of systematic structural barriers that prevent teachers from enacting practices they conceptually understand.

The qualitative data strongly support this interpretation. Teachers repeatedly described contextual constraints that made sophisticated assessment practices infeasible: " huge number of exam scripts to check for each course" (Participant 17), "extensive syllabus leading to time constraints" (Participant 29), "restrictive assessment outlines or expectations that limit the range of permissible methods" (Participant 8), and "class size, administrative responsibilities, frequent changes in institutional rules, extremely tight deadlines" (Participant 33). These are not excuses for poor practice; they are descriptions of the institutional ecology in which Bangladeshi

university teachers operate. When a lecturer “must mark 200+ examination scripts in a compressed timeframe” (Participant 8) while teaching multiple courses with minimal administrative support, performance-based assessment with detailed feedback becomes structurally impossible regardless of that teacher's conceptual knowledge.

This situated understanding aligns with DeLuca et al.'s (2019) finding that assessment literacy is "negotiated, situated, and differential across teachers and contexts" (p. 123). In their study of 453 new teachers presented with identical assessment scenarios, responses varied dramatically based on institutional policies, local cultures, and personal beliefs demonstrating that what teachers enact in practice which reflects their competence and the affordances and constraints of their specific contexts. Similarly, the present study's finding that teachers in both public and private universities showed no significant institutional differences in assessment literacy (Section 4.5.1) suggests the structural constraints. Large classes, examination mandates, limited time operate similarly across institutional types in Bangladesh, creates a relatively uniform constraint environment despite differences in resources.

5.1.2 Cultural Specificity and the Examination-Dominated Context

Brown et al.'s (2019) landmark cross-cultural factor invariance study provided critical context for interpreting this study's findings. They examined teacher conceptions of assessment across 11 jurisdictions and found that "only one model achieved configural invariance across all datasets, and even this model failed to achieve metric equivalence" (p. 16), leading them to conclude that "context, culture, and local factors shape teacher conceptions of assessment and that there is indeed no single global model." Critically, they demonstrated that teachers in examination-dominated systems (Hong Kong, Egypt, India) conceived of assessment fundamentally differently than teachers in low-stakes formative assessment environments (New Zealand,

Queensland, Cyprus).

These findings support Brown et al.'s (2019) conclusion that assessment literacy is culturally situated. The observations in this study such as moderate self-rated skills, emphasis on traditional testing, largest Skill-Use gaps in Communication of Results and Performance-Based Assessment are not deficits but logical adaptations to an examination-centric educational culture. Bangladeshi teachers' relatively strong alignment of Skill and Use in Construction of Traditional Tests (Skill M = 3.89, Use M = 3.57, gap = 0.33) compared to much larger gaps in Performance-Based Assessment (gap = 0.29) and Communication of Results (gap = 0.38) reflects a rational response to institutional expectations. Traditional exams are mandated, routinised, rewarded, regulated; whereas alternative assessments are discretionary, time-intensive, and culturally unfamiliar.

Brown et al.'s (2019) comparison of Egyptian and New Zealand teachers is particularly instructive. Egyptian teachers, operating in a system where "end-of-year exams are the sole mechanism for student progression" (Gebril & Eid, 2017, p. 89), conceived of assessment primarily as summative certification despite policy rhetoric about formative assessment. Similarly, Bangladeshi university teachers in this study described assessment practices that prioritised high-stakes midterm and final examinations, with formative practices used only "occasionally" and often constrained by institutional policies that mandate specific grade distributions from exams.

This cultural specificity challenges the applicability of Western assessment literacy frameworks to the Bangladeshi context. The API, originally developed for American K-12 teachers in formative-assessment-rich environments (Zhang & Burry-Stock, 1997), measured practices that assume certain institutional affordances such as time for rubric development, opportunities for

moderation, technological infrastructure for data analysis, cultural acceptance of teacher professional judgment. When these affordances are absent (as they largely are in Bangladeshi higher education) teachers may possess conceptual awareness of practices (reflected in moderate Skill ratings) but lack realistic opportunities for execution (reflected in lower Use ratings). This is not a failure of teachers but a mismatch between international frameworks and local assessment ecologies.

5.1.3 Domain-Specific Patterns: Explaining the Largest Gaps

The domain-level analysis revealed differential Skill-Use gaps, with Communication of Results (gap = 0.38), Construction of Traditional Tests (gap = 0.33), and Performance-Based Assessment (gap = 0.29) showing the largest discrepancies. These findings can be explained through the lens of Bangladesh's examination culture and the structural constraints teachers described.

a. Communication of Results (gap = 0.38): This domain encompasses practices such as reporting assessment outcomes to students, parents, colleagues, and administrators; explaining grades clearly; and using results to inform stakeholders about student progress. The large gap here (Skill M = 3.54, Use M = 3.17) likely reflects several contextual factors. First, in Bangladeshi higher education, "parent involvement is not typical at the tertiary level" (Section 3.3.2.1), reducing one entire category of communication practice. Second, institutional policies often prescribe rigid grading procedures, limiting teachers' autonomy in how results are communicated. Third, the qualitative data revealed that many teachers lack training in "data-based informatics for understanding teaching quality" (Participant 41) and systematic communication protocols. Finally, in an examination-dominated system, communication beyond posting final grades is often not institutionally expected or rewarded. Teachers may understand

the value of transparent, detailed communication (moderate Skill rating) but have no structural mechanisms, time, or institutional mandate to implement it (lower Use rating).

b. Performance-Based Assessment (gap = 0.29): This domain includes designing and scoring authentic tasks such as presentations, portfolios, debates, and essays with rubrics. The gap (Skill M = 3.72, Use M = 3.43) can be directly traced to the qualitative themes. Teachers explicitly mentioned that large class sizes ("huge number of students," Participant 12) and concerns about academic dishonesty ("tendencies of academic dishonesty among students," Participant 29) made performance assessments impractical. As one participant explained, "mixed ability groups" and "variability in students' prior preparation" (Participant 19) created fairness concerns when implementing open-ended tasks. Moreover, performance-based assessment requires extensive time for moderation, rubric calibration, detailed scoring. These resources are unavailable when teachers face "a huge number of exam scripts to check" (Participant 17) with "extremely tight deadlines" (Participant 33). The qualitative data thus provide a contextual explanation for a pattern that quantitative data alone might frame as simple under-utilisation. Teachers know how to design performance tasks (moderate-to-high Skill) but institutional realities prevent regular implementation (lower Use).

c. Construction of Traditional Tests (gap = 0.33): Interestingly, this domain showed both the highest absolute Skill and Use scores (M = 3.89 and 3.57, respectively) yet also one of the largest gaps. This pattern likely reflects that traditional test construction is heavily practised and institutionally mandated in Bangladesh (hence high Use relative to other domains), yet teachers still perceive a gap between what they know about quality test design and what they actually implement. The qualitative data revealed that teachers wanted training in "item analysis" and "statistical analysis of test results" (Participants 14, 41). These technical competencies would

improve traditional test quality. Thus, even in the domain most aligned with Bangladesh's examination culture, teachers recognise that their routine practice could be more rigorous if supported by training in psychometric principles.

d. Smallest gap (Selection & Planning, gap = 0.15): By contrast, Selection and Planning of assessment methods showed the smallest Skill-Use gap which suggests that planning activities (deciding which assessments to use, aligning them with objectives) translate relatively easily into practice because they occur at the design stage before time and logistical constraints bite. This finding aligns with the literature showing that "planning knowledge translated more readily into daily work" (Section 4.4.2) because it does not require sustained enactment under resource constraints.

5.2 Critical Examination of Self-Report Limitations

5.2.1 Social Desirability and Optimistic Self-Assessment

A fundamental limitation of this study is its reliance on self-report data, which are subject to well-documented biases. Teachers' self-perceived Skill ratings (M = 3.71, "moderately skilled" to "skilled") may reflect social desirability bias. This is the tendency to present oneself in a favourable light that the high Skill ratings may be partially influenced by social desirability bias, as teachers may overestimate their competence to appear professional (Paulhus & Vazire, 2007) or inflated self-assessment due to limited metacognitive awareness of one's true competence. The Dunning-Kruger effect (Kruger & Dunning, 1999) described a cognitive bias in which individuals with limited expertise overestimate their ability precisely because they lack the knowledge to recognise their deficits. In this study, 54% of participants (n = 28) reported no formal training in assessment, yet most rated themselves as moderately to highly skilled across

multiple domains. This paradox of reporting competence without formal preparation suggests that Skill ratings may reflect confidence born of necessity and routine practice rather than deep, formally validated expertise.

Qualitative responses support this interpretation. Many participants explicitly stated they had "learnt by myself" (Participant 5) or had "no formal training in assessment" (Participant 23), yet quantitative ratings rarely dipped below the midpoint of the scale. This shows that teachers may be calibrating their Skill ratings against their own limited experience base rather than against objective professional standards. A lecturer who has designed exams for three years without ever studying psychometric principles may rate themselves as "skilled" because they have developed workable routines, not because those routines would meet external quality benchmarks. As one participant reflected, "teachers' responsibility should not be solely assumed without adequate training and institutional support" (Participant 31) acknowledging a gap between felt competence and what properly supported competence would entail.

Furthermore, the remarkably high internal consistency for the overall Skill ($\alpha = .97$) and Use ($\alpha = .96$) scales, combined with low within-person variability (56% of participants had $SD < 0.5$ across all 67 Skill items), suggests possible acquiescence bias, straightlining or response sets where respondents select similar ratings across many items rather than discriminating carefully between practices (Section 4.3). While these coefficients technically indicate "excellent" reliability, values approaching 1.0 can also signal redundancy and insufficient differentiation (Tavakol & Dennick, 2011). This raises the question of whether participants were making fine-grained judgments about specific practices or applying a general, undifferentiated sense of moderate competence across the board.

5.2.2 Memory Limitations and Frequency Recall

Self-reported Use ratings are subject to distinct biases. Teachers were asked to estimate how often they performed 67 different assessment practices. It is indeed a cognitively demanding task vulnerable to recall errors and reconstruction biases (Schaeffer & Presser, 2003). Participants may have based frequency estimates on recent experiences (recency bias), on what they believe they should be doing (prescriptive bias), or on general impressions rather than accurate tallies. For example, a teacher who conducted one peer-assessment activity in the past year might rate their Use as "sometimes" (scale midpoint) because they remember doing it and perceive it as occasional practice, even though "sometimes" on a frequency scale would typically imply more regular enactment.

Moreover, the Use ratings may be sensitive to recent contextual factors. If a participant completed the survey immediately after a particularly stressful examination period, their frequency estimates might be lower due to heightened awareness of constraints. Conversely, if they had just attended a professional development workshop on formative assessment, their estimates might be optimistically inflated. The cross-sectional design (Section 3.1) means these findings represent a snapshot influenced by "temporary contextual factors such as recent policy changes or curriculum reforms" (Sedgwick, 2014) which is a limitation discussed further in Section 5.4.

5.2.3 Why Self-Report Data Remain Valuable despite Limitations

Despite these well-founded concerns, self-report data retain significant value for several reasons. First, in contexts like Bangladesh, where “little prior research exists” on university teachers’ assessment practices (Khan, 2022), even imperfect baseline data are valuable. Moderate Skill

ratings, the consistent pattern of higher self-reported Skill than frequency of Use across domains, and identified gaps in Communication and Performance-Based Assessment offer a useful starting point for hypothesis generation and professional development planning, even though the absolute values should not be interpreted literally.

Second, teachers' *perceptions* of their competence and practice are consequential in their own right. A teacher who perceives themselves as moderately skilled may approach professional development differently than one who perceives significant deficits. Similarly, teachers' awareness of Skill-Use gaps such as "I know how to do this but don't use it often" provides insight into whether barriers are primarily knowledge-based (requiring training) or structural (requiring institutional change). The dual-scale API design specifically enables this distinction, which single-scale knowledge tests cannot capture (Zhang & Burry-Stock, 2003).

Third, the convergence of quantitative and qualitative findings strengthens credibility. The largest Skill-Use gaps identified quantitatively (Communication, Performance-Based Assessment) corresponded precisely to domains where qualitative responses identified the most barriers (time constraints, large classes, lack of training). This triangulation (Creswell & Plano Clark, 2018) suggests that despite measurement imperfections, the patterns are substantively meaningful rather than purely artefactual.

Finally, self-report instruments are pragmatic and cost-effective for exploratory research with geographically dispersed populations (Kunter & Baumert, 2006). Classroom observations or performance-based assessments of teacher competence would provide more objective data but were beyond the scope of this study. The self-report approach therefore represents a methodological limitation that is acknowledged, with interpretations and justifications presented cautiously.

5.3 Integrating Qualitative and Quantitative Findings

5.3.1 Reconciling "Moderate-to-High Skill" with Extensive Reported Challenges

On the surface, there appears to be a tension between teachers' moderate-to-high self-rated Skill (overall M = 3.71, with several domains exceeding M = 3.8) and the extensive challenges they described qualitatively (lack of training, large classes, time constraints, inadequate institutional support). How can teachers claim to be "skilled" while simultaneously reporting feeling unprepared and under-resourced?

This apparent contradiction dissolves when we recognise that Skill ratings reflect self-perceived competence relative to teachers' own experience and awareness, not against objective professional standards. A teacher who has designed multiple-choice exams for five years may rate themselves as "skilled" in traditional test construction because they have developed functional routines that meet institutional expectations, not because their tests would pass external psychometric scrutiny. The qualitative responses reveal the limits of this self-taught competence such as explicit requests for training in "item analysis," "statistical analysis of test results," and "data-based informatics" (Participants 14, 38, 41) indicate awareness that current practices, while workable, lack technical sophistication.

Furthermore, the qualitative themes particularly "Lack of Assessment-Specific Training and Institutional Support" (Section 4.6.4) provide essential context for interpreting moderate Skill ratings. One participant stated: "teachers' responsibility should not be solely assumed without adequate training and institutional support" (Participant 31). Another noted: "I had no formal training...learnt by myself" (Participant 5). These comments suggest that teachers' self-rated competence is *experiential* rather than *formally validated*. They rate themselves as moderately skilled because they *are* conducting assessments (designing tests, grading scripts, assigning

marks) and those assessments function adequately within their institutional contexts. But this functional adequacy should not be confused with deep assessment literacy grounded in psychometric principles, pedagogical research, and reflective practice (DeLuca et al., 2016).

5.3.2 How Qualitative Themes Explain Quantitative Patterns

The five qualitative themes identified in Section 4.6 directly illuminate specific quantitative findings:

5.3.2.1 Theme 1: Contextual Constraints (Time, Class Size, Policy)

This theme explains why Performance-Based Assessment and Communication of Results showed large Skill-Use gaps. Teachers stated that "a huge number of exam scripts to check" (Participant 17) and "large class sizes" (Participants 12, 19, 33) made time-intensive assessment methods infeasible. When institutional policies mandate that "60% of the course grade come from a final exam" (Section 6.6), even teachers who understand performance assessment principles may default to traditional formats. The qualitative data thus reframe the quantitative gap not as individual under-utilisation but as structural constraint.

5.3.2.2 Theme 2: Student Factors (Preparedness, Engagement, Academic Integrity)

Teachers described "variability in students' prior preparation" (Participant 19), "students do not have the basic understanding" (Participant 27), and "tendencies of academic dishonesty" (Participant 29). These concerns help explain the Performance-Based Assessment gap. Teachers may feel conceptually capable of designing authentic tasks but judge them impractical when students are unmotivated or likely to cheat. As one noted, performance tasks may be "unfair in contexts where academic dishonesty is prevalent and student preparation is uneven" (Section

4.6.3). This is not deficit thinking about students but a pragmatic recognition that assessment methods must be adapted to contextual realities (Xu & Brown, 2016).

5.3.2.3 Theme 3: Lack of Training and Institutional Support

The consistent finding that 54% of participants had no formal assessment training (Table 1) contextualises moderate Skill ratings. Teachers explicitly requested training in "statistical analysis," "item analysis," "formative assessment techniques," "rubric development," and "technology-enhanced assessment" (Section 4.6.4). These requests align precisely with domains showing lower Skill scores: Standardised Testing & Data Use (M = 3.53) and Communication of Results (M = 3.54). The convergence suggests that quantitative gaps and qualitative training needs are two perspectives on the same underlying reality signifying teachers recognise areas where their self-taught competence is insufficient.

5.3.2.4 Theme 4: Tension between Traditional Exams and Alternative Assessment

Teachers described institutional pressure to adhere to "restrictive assessment outlines" and examination formats while expressing interest in "more diverse forms of assessment" (Section 4.6.5). This tension explains the finding that Construction of Traditional Tests had both high Skill (M = 3.89) and high Use (M = 3.57). Traditional exams are institutionally mandated while Performance-Based Assessment had moderate Skill (M = 3.72) but lower Use (M = 3.43) which indicates that alternative methods are discretionary and often culturally unfamiliar. The qualitative data reveal that this is not simple resistance to innovation but negotiation between pedagogical knowledge and institutional demands (Xu & Brown, 2016).

5.3.2.5 Theme 5: Technical and Digital Skills Gaps

Participants mentioned difficulties with "analysing assessment data systematically," "item analysis," "plagiarism detection tools," and "technology-enhanced assessment platforms" (Section 4.6.6). These specific technical gaps align with the finding that Standardised Testing & Data Use had the lowest overall Skill (M = 3.53) and Use (M = 3.28) scores. The qualitative data explain why. Teachers lack both conceptual knowledge and "software skills, constrained by time, or institutional infrastructure to apply them regularly" (Section 4.6.6). This finding underscores that closing assessment literacy gaps requires both training and infrastructure investment.

5.4 Comparison with International Literature

5.4.1 Parallels with European and Asian Contexts

The findings of this study align closely with international research on language teacher assessment literacy, particularly studies conducted in examination-oriented contexts or in systems with limited professional development infrastructure.

Vogt and Tsagari's (2014) survey of foreign language teachers across several European countries found that while teachers valued assessment and were comfortable with informal techniques like quizzes, many lacked training in complex tasks such as "developing tests with clear specifications or conducting item analysis" (p. 387). A significant proportion reported never having formal coursework on language assessment. This mirrors the present study, where 54% of Bangladeshi teachers had no formal training and explicitly requested training in item analysis and statistical methods (Section 4.6.6). Both studies demonstrate a common gap. Teachers develop functional assessment routines through experience but lack technical competencies that formal training would provide.

However, a key difference is that Vogt and Tsagari's (2014) participants worked in diverse European systems with varying degrees of high-stakes testing, whereas Bangladeshi teachers

operate uniformly within an examination-centric culture. This contextual difference explains why Bangladeshi teachers showed higher Use in Construction of Traditional Tests relative to other domains, Traditional exams are not merely one option but the institutional default.

The study of English teachers in China by Gan et al. (2004) found that teachers "felt uncertain about designing performance-based assessments and using scoring rubrics due to minimal training opportunities" (p. 412). Teachers reported avoiding certain assessment types (e.g., oral assessments) because scoring was perceived as "difficult" and "time-consuming" (p. 415). This construct is virtually identical to the present study's findings. Bangladeshi teachers showed one of the largest Skill-Use gaps in Performance-Based Assessment (gap = 0.29) and qualitatively described concerns about large classes, academic dishonesty, fairness, and equity when implementing open-ended tasks (Section 4.6.3). Both studies suggest that in resource-constrained Asian higher education systems, teachers may possess awareness of performance assessment principles but lack institutional support to implement them reliably.

The similarity of findings across Bangladesh, China, and Europe despite different educational systems suggests that certain patterns are relatively robust: (1) teachers often lack formal assessment training, (2) time and class size are universal constraints on formative and performance-based assessment, and (3) technical domains (item analysis, data use) are consistently identified as weak areas requiring professional development.

5.4.2 What Is Distinctive About the Bangladeshi Context?

While parallels exist with international studies, several findings are distinctive to (or at least particularly pronounced in) the Bangladeshi context:

a. Examination dominance: The strength of alignment between Skill and Use in Construction of Traditional Tests (both scores among the highest across all domains) reflects Bangladesh's

particularly rigid examination culture. Unlike European systems where teachers often have autonomy to balance formative and summative assessment, or even East Asian systems where performance tasks are increasingly incorporated alongside exams (Gan et al., 2018), Bangladeshi university policies often mandate that 60%+ of grades come from midterm and final exams (Section 6.6). This institutional rigidity narrows the space for enacting alternative assessment regardless of teacher competence.

b. Absence of empirical precedent: A distinctive feature of this study is that it provides the *first* quantitative profile of assessment literacy among Bangladeshi higher education English teachers. As noted in Section 2.5.3, no published research in Bangladesh has quantitatively measured university English teachers' self-reported skill and use profiles across key domains of assessment literacy. Prior research consisted of small-scale qualitative studies or secondary school samples. This study thus fills a critical empirical gap and establishes baseline data against which future research and interventions can be compared.

c. Language teaching specificity: Unlike general teacher assessment literacy studies, this research focuses specifically on English language teachers, who face unique challenges in assessing communicative competence, speaking and writing performance, and language proficiency development (Inbar-Lourie, 2008). The finding that Communication of Results had the largest Skill-Use gap may partly reflect the complexity of explaining language assessment results which often involve subjective judgments about fluency, coherence, and communicative effectiveness compared to clearer-cut content subjects. This language-specific dimension warrants further exploration in future research.

5.4.3 Implications for Universality of Assessment Literacy Frameworks

The findings of this study, interpreted through Brown et al.'s (2019) cross-cultural factor invariance research, challenge assumptions that Western-developed assessment literacy frameworks can be straightforwardly applied in non-Western contexts. Brown et al. demonstrated that assessment conceptions vary systematically with national policy contexts, examination cultures, and educational values. The present study provides empirical evidence that supports this cultural specificity in a previously under-researched context.

Specifically, this study found that:

1. Traditional test construction skills and practices were relatively well-developed and aligned (both Skill and Use > 3.5), reflecting institutional priorities and mandates.
2. Performance-based assessment and communication practices showed large gaps, reflecting that these domains are less culturally embedded and institutionally supported in Bangladesh.
3. No significant differences emerged between public and private universities, suggesting that structural constraints operate uniformly across institutional types. Such finding may be specific to Bangladesh's relatively centralised higher education governance.

These patterns would likely not emerge in New Zealand or Australian university samples, where performance-based assessment and transparent communication are culturally normalised and institutionally expected (Brown, 2004). Conversely, a New Zealand-developed assessment literacy framework emphasising formative feedback, student self-assessment, and collaborative moderation (Xu & Brown, 2016) may *misalign* with Bangladeshi teachers' needs precisely because it assumes institutional affordances and cultural norms that do not exist.

This does not mean Bangladeshi teachers are "less assessment literate" than Western counterparts. Rather, it means they have developed a form of assessment literacy *adapted to their context* that prioritises examination integrity and efficient marking of large cohorts alongside navigating institutional constraints. From this perspective, asking Bangladeshi teachers to adopt Western formative assessment models without addressing structural barriers (class sizes, time, policy mandates) is akin to asking a teacher in a low-resource setting to implement technology-enhanced assessment without providing computers. The problem is not individual competence but systemic mismatch.

Future professional development and policy interventions must therefore be designed with explicit attention to Bangladesh's assessment ecology rather than assuming that "international best practices" can be imported wholesale. This might mean, for example, focusing on how to improve the quality of traditional tests (which teachers *do* use regularly) through better item-writing and analysis techniques, rather than emphasising performance-based assessment that current institutional realities make infeasible for most teachers.

5.5 Factors Influencing Assessment Literacy: Experience, Training, and Institution Type

5.5.1 Teaching Experience as a Predictor of Assessment Literacy

Spearman's rho correlations revealed significant positive relationships between years of teaching experience and several assessment literacy domains (Section 4.5.3). More experienced teachers reported substantially higher self-perceived skills in Grading Practices ($\rho = .482$, $p < .001$) and Performance-Based Assessment ($\rho = .467$, $p < .001$), with moderate correlations for Use of Assessment Results [Skills ($\rho = .405$, $p = .003$)] and Performance-Based Assessment [Use ($\rho = .403$, $p = .003$)]. Overall assessment literacy correlated significantly with experience for both

Skill ($\rho = .447$) and Use ($\rho = .353$), both $p < .01$.

These findings align with existing research showing that assessment literacy "accumulates through both training and on-the-job experience" (Stiggins, 1999; Giraldo, 2018). However, a critical nuance emerged. Experience correlated more strongly with Skill perceptions than with Use frequency (e.g., Grading Practices: Skill $\rho = .482$ vs. Use $\rho = .365$). This divergence suggests that veteran teachers feel increasingly confident about assessment practices as they gain experience, but this growing confidence does not translate proportionally into more frequent enactment. In other words, experience contributes more to teachers' self-perceived competence than to changes in their day-to-day assessment behaviour.

This finding reinforces the TALiP framework's emphasis on contextual constraints (Xu & Brown, 2016). Even as teachers accumulate years of experience and develop greater assessment knowledge, the structural barriers they face (large classes, time pressures, institutional policies) remain relatively constant. Thus, a teacher with 15 years of experience may feel *much* more skilled than a novice (reflected in Skill ratings) but still use performance-based assessment only occasionally (reflected in Use ratings) because the constraints preventing enactment have not changed.

Notably, two domains showed no significant correlation with experience: Traditional Test Construction ($\rho = .21$, $p = .13$) and Interpretation of Standardised Tests ($\rho = .22$, $p = .11$). This suggests that technical, psychometric domains do not improve simply through years of practice, they require formal training. Teachers may write dozens of exams over a career without developing deeper understanding of item analysis, reliability, or validity unless they receive explicit instruction in these concepts. This finding has direct implications for professional development priorities. Technical skills must be *taught*, not assumed to develop organically

through experience.

5.5.2 Formal Assessment Training: A Modest but Consistent Advantage

Comparisons of teachers with formal assessment training (n = 24) versus those without (n = 28) revealed that trained teachers scored higher across all domains, with effect sizes consistently in the small-to-medium range (Cliff's δ = –0.28 to –0.34, all negative indicating higher trained group scores). However, after controlling for false discovery rate, none of these differences reached statistical significance (Section 4.5.2).

Such consistent trends that fail to reach statistical significance likely reflects limited statistical power due to small sample size (N = 52) combined with genuine but modest effects. The qualitative finding that many trained teachers described their training as brief ("only briefly covered" in general pedagogy courses, Participant 15) suggests that most had not received intensive, sustained assessment education. Thus, while training confers advantages, these advantages are modest when training itself is superficial.

Nonetheless, the consistent directionality across all domains is theoretically meaningful. It suggests that even limited formal training provides teachers with frameworks, vocabulary, and awareness that shape their assessment thinking, even if effects on practice remain modest when institutional constraints dominate. The finding that nominally significant pre-correction differences appeared in Grading Practices (p = .049) and Use of Assessment Results (p = .038) suggests these may be domains where formal training has the most tangible impact, perhaps because they involve applying systematic frameworks (rubrics, criteria, data interpretation) that training can efficiently convey.

5.5.3 No Institutional Differences: Uniform Constraints across Public and Private Universities

Contrary to expectations, Mann-Whitney U tests revealed no significant differences in assessment literacy between public university (n = 19) and private university (n = 33) teachers across any domain (Section 4.5.1). This null finding is itself informative. It suggests that the structural constraints shaping assessment practice operate relatively uniformly across institutional types in Bangladesh.

While public and private universities differ in resources, governance structures, and student populations (Alam et al., 2020), these differences do not translate into divergent assessment practices. This may reflect several contextual realities:

1. **Uniform policy mandates:** National accreditation requirements and degree standards may impose similar examination formats and grading structures across all universities, regardless of institutional type.

2. **Shared cultural norms:** The examination-centric culture pervades Bangladesh's entire educational system (Sultana, 2019). Even well-resourced private universities default to traditional assessment formats because that is what students, parents, and employers expect and understand.

3. **Similar constraint profiles:** While public universities are often characterised by very large classes (sometimes 100+ students), private universities, despite smaller classes on average face other constraints such as heavy teaching loads, limited administrative support, and commercial pressures to maintain high pass rates. The net effect is that teachers in both sectors face significant barriers to implementing time-intensive formative or performance-based assessment.

This finding cautions against simplistic assumptions that improving assessment practices is primarily a resourcing issue. Simply providing private-university-level resources to public

universities or vice-versa may not substantially change assessment practices if cultural norms, policy mandates, and institutional inertia remain unaddressed.

Chapter 6: Study Limitations

6.1 Self-Report Data and Measurement Bias

This study's primary limitation is its reliance on self-reported assessments of teacher competence and practice frequency, which are subject to multiple biases. The API measures teachers' perceived skills and self-reported frequency of use, not objectively assessed competence or directly observed classroom behaviour. Teachers' Skill ratings may be inflated by social desirability bias (the tendency to present oneself in a favourable light) or by the Dunning-Kruger effect, whereby individuals with limited expertise lack the metacognitive frameworks to accurately evaluate their own competence. Kruger and Dunning (1999) demonstrated that individuals with low knowledge often overestimate their abilities because they lack the expertise necessary to recognise gaps in their understanding. Teachers in this study who reported no formal assessment training (54% of the sample) may rate themselves as "moderately skilled" based on developed routines or functional exam-writing experience, yet lack systematic frameworks that would reveal technical gaps such as item bias, unreliable grading procedures, or inappropriate test specifications.

Similarly, Use ratings are vulnerable to recall limitations and reconstruction bias. Teachers asked to estimate how often they performed 67 different practices may have relied on recent experiences, aspirational beliefs about what they should be doing, or general impressions rather than accurate tallies of actual behaviour. Additionally, what teachers report as their practice at the time of survey completion may differ from their sustained practices over an academic year. Cross-sectional snapshots are particularly susceptible to temporary contextual influences such as proximity to examination periods or recent curriculum changes.

However, self-report data retain considerable value for this study. First, in contexts where little prior empirical research exists, even imperfect self-report data provide essential baseline information for professional development planning. Second, teachers' perceptions of their skills and practice gaps are consequential in their own right, as they influence motivation for professional learning and self-directed improvement. Third, the convergence between quantitative patterns and qualitative explanations strengthens confidence in the findings. The largest Skill-Use gaps appeared precisely in domains where teachers qualitatively described structural barriers (Communication of Results, Performance-Based Assessment). This alignment between methods suggests the patterns reflect genuine features of practice environments rather than measurement artefacts alone.

6.2 Instrument Design and Contemporary Framework Mismatch

The API was originally developed for American K-12 teachers and reflects traditional psychometric conceptualisations of assessment literacy. While adapted for the Bangladeshi higher education context, the instrument emphasises test construction, grading, and standardised test interpretation rather than capturing contemporary emphases on assessment for learning, student self-assessment, peer assessment, and assessment as a situated social practice embedded in disciplinary contexts. The literature review presented a comprehensive, contemporary view of assessment literacy as contextually-situated and developmental. The API, by contrast, reflects a narrower measurement-oriented framework.

This mismatch is acknowledged as deliberate and pragmatically justified. No single validated instrument fully captures contemporary situated frameworks in a way that is practical for large-scale survey research. Developing a new instrument was beyond this study's scope. Despite its traditional focus, the API covers foundational practices shown to matter for student learning such

as test quality, performance assessment, grading, feedback, and communication. These practices remain meaningful in Bangladesh's exam-oriented context. Additionally, the qualitative component was designed to capture contemporary concerns about student engagement, formative feedback, and assessment for learning that the quantitative items alone might miss.

The instrument length (67 items with dual ratings) also presented challenges. Item count may have induced respondent fatigue, potentially leading to satisficing behaviour where teachers provided similar ratings across many items rather than carefully discriminating between practices. Evidence of this appears in the very high overall Cronbach's alphas ($\alpha = 0.97$ for Skill, 0.96 for Use) combined with 56% of participants showing low within-person variability (standard deviation below 0.5), suggesting possible response similarity rather than genuine consistency.

6.3 Sampling Strategy and Generalisability

The study employed purposive and convenience sampling. Purposive sampling ensured participants had direct experience teaching undergraduate English essential for relevance to research questions. Convenience sampling through accessible networks (department contacts, professional associations, social media groups) facilitated practical recruitment but limits generalisability. The sample was not randomly selected and cannot be assumed representative of all Bangladeshi university English teachers.

Several characteristics constrain generalisability. First, geographic concentration was visible. 90% of participants ($n = 47$) were based in Dhaka, the capital city where institutional resources and teaching conditions may differ substantially from provincial universities. Second, career stage skew was found. The sample was predominantly early-career, with median experience of

2.25 years and 50% having 1.48 years or less. Findings therefore primarily reflect assessment practices among novice teachers learning institutional norms, not across the profession. Senior faculty with refined approaches over decades may have different skill profiles and practice patterns. Third, self-selection bias was prominent. Teachers who were more accessible through professional networks, more active in assessment discussions, or more interested in professional development may have been overrepresented. Teachers less engaged with the profession or less willing to reflect on their practices may be underrepresented.

Fourth, sample size (N = 52) was modest. While adequate for descriptive purposes, it limited statistical power to detect small-to-medium effects in group comparisons. Non-significant findings for training effects or institutional differences may reflect insufficient power rather than genuine null effects. Confidence intervals and effect size estimates are therefore more informative than p-values alone.

Finally, convenience samples are particularly vulnerable to unmeasured confounding. Teachers who volunteered may differ systematically on variables not captured (motivation, professional identity, workload satisfaction) that could influence both assessment literacy and willingness to participate.

6.4 Cross-Sectional Design and Causal Inference

This study collected data at a single time point, providing a snapshot of current practices rather than tracking development over time. Cross-sectional designs cannot establish causal relationships. While experience correlated with assessment literacy, particularly for Grading Practices and Performance-Based Assessment, the direction of causation cannot be determined. Does experience build assessment competence, or do more assessment-literate teachers simply

remain in the profession? The findings represent associations, not evidence of development through experience.

Similarly, the study cannot determine whether professional development interventions would close observed Skill-Use gaps or produce changes in practice. Contextual factors may prevent enactment regardless of teacher knowledge. Establishing causation would require longitudinal designs tracking the same teachers over years or intervention studies experimentally manipulating professional development or contextual conditions.

The cross-sectional snapshot is also sensitive to temporal context. Findings may be influenced by recent policy changes, examination cycles, or curriculum reforms. Conclusions drawn from data collected in one semester may not generalise to other academic years.

6.5 Measurement Assumptions: Skill-Use Comparison

The study compared Skill and Use ratings despite their different scale anchors. Skill reflected competence judgements (1 = not at all skilled to 5 = highly skilled), while Use reflected frequency (1 = never to 5 = very often). These are conceptually distinct constructs measured on different implicit reference frames. A teacher rating themselves as "4 = skilled" and "3 = sometimes use" does not necessarily indicate an absolute gap on a shared metric. Rather, it indicates that self-perceived capability and self-reported enactment frequency are related but distinct dimensions.

The Wilcoxon tests identified whether one self-report dimension (Skill) tended to be rated higher than the other (Use), but cannot definitively explain the gap. Multiple interpretations remain plausible. Skill ratings may capture perceived potential under ideal conditions, while Use ratings reflect practice under real constraints. Alternatively, Skill may reflect what teachers have learned

to do, while Use reflects what institutional systems allow or encourage. Social desirability might inflate Skill more than Use if teachers believe admitting high capability looks better than admitting low frequency.

Therefore, the observed Skill-Use differences are patterns in self-report rather than evidence of unused knowledge or direct measures of implementation barriers.

6.6 Contextual Constraints beyond Individual Measurement

The study did not systematically measure institutional policies, resource availability, class sizes, or workload, the contextual factors that qualitative responses identified as powerful determinants of assessment practice. While qualitative themes described large classes (often exceeding 100 students), time constraints, and restrictive examination mandates, the quantitative analysis lacked variables to directly model these constraints. Consequently, the study cannot precisely estimate how much of the observed Skill-Use gap reflects individual knowledge deficits versus structural barriers.

Teachers' assessment practices are shaped by policies they do not control. Many described mandatory examination formats, fixed assessment weights (often 60% or more exam-based), and centralised marking schemes that limit discretionary practice. These constraints operate uniformly across the sample regardless of individual competence, potentially masking variation in actual capability. The finding that public and private universities showed no significant differences despite private institutions' presumed resource advantages suggests that uniform policy mandates and examination cultures outweigh institutional type as determinants of practice. Similarly, the study did not measure class sizes, marking loads, or time available for assessment design. These factors were repeatedly cited in qualitative responses as barriers.

Incorporating these variables could substantially reframe interpretation of the Skill-Use patterns to "systemic constraints prevent enactment regardless of knowledge."

Chapter 7: Conclusion

7.1 Summary of Key Findings

This study profiled the assessment literacy of 52 undergraduate English language teachers in Bangladesh, revealing consistent patterns across quantitative and qualitative data. Teachers reported moderate self-perceived skill (M = 3.71) consistently exceeding self-reported practice frequency (M = 3.44) across all seven assessment domains. Largest gaps appeared in Communication of Results (Mdiff = 0.38), Construction of Traditional Tests (Mdiff = 0.33), and Performance-Based Assessment (Mdiff = 0.29). These are the domains that require time and collaboration in an institutional infrastructure. Traditional test construction showed strongest Skill-Use alignment, reflecting alignment with exam-centric institutional culture.

Teaching experience correlated significantly with assessment literacy, particularly for Grading Practices and Performance-Based Assessment. Teachers with formal assessment training scored consistently higher across domains, though differences did not reach statistical significance after controlling for multiple testing, likely reflecting limited sample size and training intensity.

Qualitatively, five themes explained quantitative patterns. Contextual constraints (large classes, heavy marking, restrictive policies) limited implementation of time-intensive practices. Student factors (variable preparation, academic dishonesty) discouraged open-ended assessment formats. Lack of formal training, particularly in technical domains like item analysis and data interpretation, reflected gaps in institutional professional development. Teachers expressed tension between recognising value in formative and performance-based assessment and operating within exam-oriented institutional cultures. Finally, technical and digital skills gaps in data analysis limited sophisticated use of assessment evidence.

These findings support Brown et al. (2019) conclusion that assessment literacy is culturally situated. What constitutes assessment-literate practice varies with examination dominance, resource constraints along with institutional expectations. Bangladeshi teachers have developed assessment competence adapted to their institutional realities rather than evidence of individual deficits.

7.2 Theoretical Contributions

The study operationalised Xu and Brown (2016) Teacher Assessment Literacy in Practice (TALiP) framework, demonstrating that Skill-Use gaps reflect negotiated compromises between knowledge and context rather than simple incompetence. Teachers continuously balance assessment conceptions against institutional constraints and policy mandates. The gaps observed represent settlements in this negotiation rather than unused capacity. This reframes responsibility for improving assessment from individual teachers to institutional and policy systems that must provide enabling conditions such as time, infrastructure, training, autonomy, and cultural legitimacy for diverse assessment approaches.

The dual-scale design proved valuable as a heuristic for profiling assessment practice. While imperfect, comparing Skill and Use ratings on different anchors identified areas where perceived competence exceeds reported practice, pointing to potential intervention targets. This approach could inform professional development prioritisation in contexts where resources are limited.

The study's largest contribution is empirical. It provides the first quantitative profile of assessment literacy in Bangladeshi higher education, an underresearched context where examination cultures, resource constraints, and limited assessment training shape practices distinctly from Western contexts. This baseline enables comparison with other examination-dominated systems and demonstrates that universal frameworks require adaptation.

7.3 Professional Development Recommendations

Effective professional development must be targeted, contextualised, and should address structural barriers alongside individual capacity. Priority should focus on domains showing largest Skill-Use gaps and lowest absolute skill scores. Communication of Results and Standardised Testing-Data Use warrant immediate attention. These domains require technical knowledge rarely taught informally. Professional development should emphasise item analysis, basic statistics, communication protocols, and feedback mechanism adapted to Bangladesh's assessment environment rather than importing Western frameworks assuming different institutional affordances.

Performance-based assessment requires scalable strategies acknowledging large class constraints. Peer assessment proves effective for scaling authentic assessment in resource-limited settings, with evidence showing students benefit from evaluative engagement itself (Kulkarni et al., 2013; Patchan et al., 2017). Professional development should emphasize peer-marked formative tasks, student self-assessment protocols, and portfolio sampling manageable with 80–100 students. Sustained communities of practice within departments nurture deeper learning than isolated workshops (Patton & Parker, 2017; Vescio et al., 2008). Critically, professional development must acknowledge structural constraints rather than framing gaps as individual deficits, as deficit framing demoralises teachers facing genuine systemic barriers (Kennedy, 2014). Effective improvement requires both individual learning and institutional reform.

7.4 Institutional Policy Recommendations

Universities should revise examination mandates to allocate at least 40% of grades to non-exam assessment, creating legitimate space for formative practices without overwhelming workload (Wiliam, 2011). Where classroom sizes exceed 80 students, teaching assistants or part-time

graders provide critical support for detailed feedback (Jones, 2020). Learning management systems enabling online assignment submission, plagiarism detection, and rubric-based marking streamline large cohort assessment (Sanchez et al., 2024). Reducing class sizes strategically in skills courses (speaking, writing) where individual feedback matters most would address the single most frequently cited barrier (De Paola et al., 2013; Jones, 2020). Recognising assessment innovation in promotion criteria aligns institutional incentives with professional development (Lim et al., 2025). Creating assessment committees within departments to develop shared rubrics, conduct moderation, and review test quality builds communities of practice while distributing workload (Tummons, 2024).

7.5 National Policy Directions

National-level initiatives should establish assessment competency standards for university faculty, similar to existing teaching qualification requirements. A National Centre for Assessment in Higher Education could provide systematic capacity building, develop locally-appropriate frameworks rather than assuming imported models, and disseminate evidence about effective assessment practices in Bangladesh's context.

Policy should explicitly balance examination requirements with formative assessment opportunities. While national examinations serve accountability functions, their dominance in institutional assessment policies has narrowed the assessment repertoire. Creating policy space for alternative assessment through revised accreditation standards or examination weighting adjustments could support institutional innovation without dismantling accountability systems.

7.6 Future Research Directions

This study's limitations point to necessary research directions. Longitudinal studies tracking

cohorts of teachers over 2–3 years would clarify whether experience develops assessment literacy, whether professional development produces sustained change in practice, and how teachers' assessment identities evolve. Observational studies validating self-reports against classroom practice would estimate bias magnitude and identify where perceptions diverge from behaviour.

Comparative research across examination-dominated and formative-assessment-rich contexts would test whether observed patterns (large Skill-Use gaps, emphasis on traditional testing) are specific to Bangladesh or reflect broader examination culture effects. Developing Bangladesh-specific assessment literacy instruments would provide valid measurement aligned with local practices and institutional values rather than relying on Western-developed tools requiring extensive adaptation.

Investigation of why public and private institutions showed no differences despite presumed resource gaps would illuminate how uniform policy mandates and examination culture override institutional type. Finally, participatory action research is needed where teachers collaboratively design, implement, evaluate, and execute professional development innovations. This could generate practitioner-based evidence about what reforms work within Bangladesh's authentic constraints rather than in idealised conditions. These recommendations assume that professional development and institutional support are implemented in tandem. Professional development alone, without addressing structural barriers such as class size, time constraints, and examination-mandated policies, is unlikely to substantially shift assessment practice. Similarly, policy changes without accompanying teacher development may encounter resistance. Both must be addressed simultaneously.

References

AFT, NCME, & NEA. (1990). Standards for teacher competence in educational assessment of students. Educational Measurement: Issues and Practice, 9(4), 30--32. https://doi.org/10.1111/j.1745-3992.1990.tb00391.x

Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179–211.

Alam, G. M., Forhad, A. R., & Ismail, I. A. (2020). Can education as an “international commodity” be the backbone or cane of a nation in the era of the fourth industrial revolution? A comparative study. Technological Forecasting and Social Change, 159, Article 120184. https://doi.org/10.1016/j.techfore.2020.120184

Alkharusi, H. A. (2011). An analysis of the internal and external structure of the Teacher Assessment Literacy Questionnaire. International Journal of Learning, 18(1), 515–528. https://doi.org/10.18848/1447-9494/cgp/v18i01/47461

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). American Psychological Association.

Arbuckle, J. L. (2019). Amos (Version 26.0) [Computer software]. IBM SPSS.

Assessment Reform Group. (2002). Assessment for learning: 10 principles. University of Cambridge Faculty of Education.

Baker, B., & Taylor, L. (Eds.). (2024a). Language assessment literacy and competence, Volume 1: Research and reflections from the field (Studies in Language Testing, 55). Cambridge

University Press & Assessment.

Baker, B., & Taylor, L. (Eds.). (2024b). Language assessment literacy and competence, Volume 2: Case studies from around the world (Studies in Language Testing, 56). Cambridge University Press & Assessment.

Bartlett, M. S. (1954). A note on the multiplying factors for various chi square approximations. Journal of the Royal Statistical Society, Series B, 16(2), 296--298.

Benjamini, Y. (2010). Discovering the false discovery rate. Journal of the Royal Statistical Society: Series B, 72(2), 405–416.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289–300.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Biggs, J. B. (1996). Western misperceptions of the Confucian-heritage learning culture. In D. A. Watkins & J. B. Biggs (Eds.), The Chinese learner: Cultural, psychological and contextual influences (pp. 45--67). Comparative Education Research Centre & Australian Council for Educational Research.

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

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. Higher Education, 18(5), 529--549. https://doi.org/10.1007/BF00138746

British Educational Research Association. (2018). Ethical guidelines for educational research (4th ed.). BERA.

Brookhart, S. M. (1999). The art and science of classroom assessment: The missing part of pedagogy. ASHE-ERIC Higher Education Report, 27(1), 1–139.

Brookhart, S. M. (2008). How to give effective feedback to your students. Association for Supervision and Curriculum Development.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. Educational Measurement: Issues and Practice, 30(1), 3–12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Brookhart, S. M. (2013). The use of teacher judgement for summative assessment in the USA. Assessment in Education: Principles, Policy & Practice, 20(1), 69–90. https://doi.org/10.1080/0969594X.2012.703170

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. Assessment in Education: Principles, Policy & Practice, 11(3), 301–318. https://doi.org/10.1080/0969594042000304609

Brown, G. T. L., & Remesal, A. (2012). Prospective teachers' conceptions of assessment: A cross-cultural comparison. The Spanish Journal of Psychology, 15(1), 75--89.

https://doi.org/10.5209/rev_SJOP.2012.v15.n1.37286

Brown, G. T. L., & Remesal, A. (2017). Teachers' conceptions of assessment: Comparing two inventories with Ecuadorian teachers. Studies in Educational Evaluation, 52, 48--58. https://doi.org/10.1016/j.stueduc.2016.12.001

Brown, G. T. L., Gebril, A., & Michaelides, M. P. (2019). Teachers' conceptions of assessment: A global phenomenon or a global localism. Frontiers in Education, 4(16). https://doi.org/10.3389/feduc.2019.00016

Brown, T. A. (2015). Confirmatory factor analysis for applied research (2nd ed.). The Guilford Press.

Byrne, B. M. (2010). Structural equation modeling with AMOS: Basic concepts, applications, and programming (2nd ed.). Routledge.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56(2), 81--105. https://doi.org/10.1037/h0046016

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Cohen, L., Manion, L., & Morrison, K. (2018). Research methods in education (8th ed.). Routledge.

Conover, W. J. (1999). Practical nonparametric statistics (3rd ed.). John Wiley & Sons.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE Publications.

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). Sage.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297--334. https://doi.org/10.1007/BF02310555

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. Assessing Writing, 28, 43–56. https://doi.org/10.1016/j.asw.2016.03.001

Darling-Hammond, L. (2006). Constructing 21st-century teacher education. Journal of Teacher Education, 57(3), 300--314. https://doi.org/10.1177/0022487105285962

Das, S., Shaheen, R., Shrestha, P., Rahman, A., & Khan, R. (2014). Policy versus ground reality: Secondary English language assessment system in Bangladesh. Curriculum Journal, 25(3), 326–343. https://doi.org/10.1080/09585176.2014.909323

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

De Paola, M., Ponzo, M., & Scoppa, V. (2013). Class size effects on student achievement:

heterogeneity across abilities and fields. Education Economics, 21(2), 135–153. https://doi.org/10.1080/09645292.2010.511811

DeLuca, C., & Bellara, A. (2013). The current state of assessment education: Aligning policy, standards, and teacher education curriculum. Journal of Teacher Education, 64(4), 356–372. https://doi.org/10.1177/0022487113488144

DeLuca, C., Coombs, A., LaPointe-McEwan, D., & Chalas, A. (2019). Toward a differential and situated view of assessment literacy: Studying teachers' responses to classroom assessment scenarios. Frontiers in Education, 4(94). https://doi.org/10.3389/feduc.2019.00094

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. Educational Assessment, Evaluation and Accountability, 28(3), 251--272. https://doi.org/10.1007/s11092-015-9233-6

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. Journal of Organizational Behavior, 36(2), 171–181. https://doi.org/10.1002/job.1962

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, phone, mail, and mixed-mode surveys: The tailored design method (4th ed.). Wiley.

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

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health,

education, and the workplace. Psychological Science in the Public Interest, 5(3), 69--106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. Journal of Advanced Nursing, 62(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Elsevier. (2020). Mendeley Desktop (Version 1.19.8) [Computer software]. Elsevier.

Field, A. (2018). Discovering statistics using IBM SPSS statistics: And sex and drugs and rock 'n' roll (5th ed.). SAGE Publications.

Fowler, F. J. (2014). Survey research methods (5th ed.). SAGE Publications.

Fulcher, G. (2012). Assessment literacy for the language classroom. Language Assessment Quarterly, 9(2), 113--132. https://doi.org/10.1080/15434303.2011.642041

Gan, Z., Humphreys, G., & Hamp-Lyons, L. (2004). Understanding successful and unsuccessful EFL students in Chinese universities. *The Modern Language Journal, 88*(2), 229--244. https://doi.org/10.1111/j.0026-7902.2004.00227.x

Gan, Z., Leung, C., He, J., & Nang, H.-H. (2018). Classroom assessment practices and learning motivation: A case study of Chinese EFL students. TESOL Quarterly, 52(2), 408--438. https://doi.org/10.1002/tesq.476

Gebril, A., & Eid, M. (2017). Test preparation beliefs and practices in a high-stakes context: A teacher perspective. In D. Tsagari & J. Banerjee (Eds.), Handbook of second language assessment (pp. 263--280). De Gruyter.

George, D., & Mallery, P. (2016). IBM SPSS statistics 23 step by step: A simple guide and reference (14th ed.). Routledge.

Giraldo, F. (2018). Language assessment literacy: Implications for language teachers. Profile: Issues in Teachers' Professional Development, 20(1), 179–195. https://doi.org/10.15446/profile.v20n1.62089

Giraldo, F. (2021). Language assessment literacy: Implications for language teachers. Profile: Issues in Teachers' Professional Development, 23(1), 265--278. https://doi.org/10.15446/profile.v23n1.87149

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. Educational Measurement: Issues and Practice, 33(2), 14--18. https://doi.org/10.1111/emip.12030

Gravetter, F. J., & Wallnau, L. B. (2017). Statistics for the behavioral sciences (10th ed.). Cengage Learning.

Guskey, T. R. (2003). How classroom assessments improve learning. Educational Leadership, 60(5), 6–11.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis (7th ed.). Prentice Hall.

Harlen, W. (2006). The role of assessment in developing motivation for learning. In J. Gardner (Ed.), Assessment and learning (pp. 61–80). SAGE. DOI:10.4135/9781446250808.n11

Hattie, J. (2008). Visible learning: A synthesis of over 800 meta-analyses relating to

achievement. Routledge. https://doi.org/10.4324/9780203887332

Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions and organizations across nations (2nd ed.). SAGE Publications.

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative Health Research, 15(9), 1277–1288. https://doi.org/10.1177/1049732305276687. https://doi.org/10.1002/tesq.180

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1--55. https://doi.org/10.1080/10705519909540118

Hunter, D., & Zaman, T. (2022). English language teaching, learning and assessment in Bangladesh: Policies and practices in the school education system. British Council. https://www.teachingenglish.org.uk/sites/teacheng/files/2022-06/ELT_learning_assessment_Bangladesh_June_2022.pdf

IBM Corp. (2019). IBM SPSS Statistics for Windows (Version 26.0) [Computer software]. IBM Corp.

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

Inbar-Lourie, O. (2013). Guest editorial to the special issue on language assessment literacy. Language Testing, 30(3), 301--307. https://doi.org/10.1177/0265532213480126

Jamieson, S. (2004). Likert scales: How to (ab)use them. Medical Education, 38(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Jones, W. (2020). Does size really matter in university preparatory English language classes? Issues in Educational Research, 30(1), 135–152.

Kaiser, H. F. (1974). An index of factorial simplicity. Psychometrika, 39(1), 31--36. https://doi.org/10.1007/BF02291575

Kaiser, K. (2009). Protecting respondent confidentiality in qualitative research. Qualitative Health Research, 19(11), 1632–1641. https://doi.org/10.1177/1049732309350879

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1--73. https://doi.org/10.1111/jedm.12000

Kennedy, A. (2014). Understanding continuing professional development: The need for theory to impact on policy and practice. Professional Development in Education, 40(5), 688–697. https://doi.org/10.1080/19415257.2014.955122

Khan, R. (2022). Exploring Assessment Literacy of Tertiary-Level Teachers in Bangladesh. In R. Khan, A. Bashir, B. L. Basu, & M. E. Uddin (Eds.), Local Research and Glocal Perspectives in English Language Teaching (pp. 345-361). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-6458-9_22

Kim, Y., Dykema, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. Social Science Computer Review, 37(2), 212–233.

https://doi.org/10.1177/0894439317752406

Kline, R. B. (2016). Principles and practice of structural equation modeling (4th ed.). The Guilford Press.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology, 5(3), 213–236. https://doi.org/10.1002/acp.2350050305

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6), 1121--1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., & Klemmer, S. R. (2013). Peer and self assessment in massive online classes. In Proceedings of the 2013 Conference on Computer Supported Cooperative Work (pp. 113–122). ACM.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. Learning Environments Research, 9(3), 231–251. https://doi.org/10.1007/s10984-006-9015-7

Lavrakas, P. J. (2008). Encyclopedia of survey research methods (Vols. 1–2). Sage.

Lim, B.H., D'Ippoliti, C., Dominik, M. et al. Regional and institutional trends in assessment for academic promotion. Nature 638, 459–468 (2025). https://doi.org/10.1038/s41586-024-08422-9

Linn, R. L., & Gronlund, N. E. (2000). Measurement and assessment in teaching (8th ed.).

Prentice Hall.

Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). Wiley.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Addison-Wesley.

MacLellan, E. (2004). How convincing is alternative assessment for use in higher education? Assessment & Evaluation in Higher Education, 29(3), 311–321.

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. Language Testing, 30(3), 329--344. https://doi.org/10.1177/0265532213480129

Mann, C. J. (2003). Observational research methods. Research design II: Cohort, cross sectional, and case-control studies. Emergency Medicine Journal, 20(1), 54–60. https://doi.org/10.1136/emj.20.1.54

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics, 18(1), 50–60. https://doi.org/10.1214/aoms/1177730491

McDonald, R. P. (1999). Test theory: A unified treatment. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410601087

Melissa M. Patchan, Christian D. Schunn & Russell J. Clark (2017): Accountability in peer assessment: examining the effects of reviewing grades on peer ratings andpeer feedback, Studies in Higher Education, DOI: 10.1080/03075079.2017.1320374. http://dx.doi.org/10.1080/03075079.2017.1320374

Menken, K., Hudson, T., & Leung, C. (2014). Symposium: Language assessment in

Mertler, C. A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? American Secondary Education, 33(1), 49–64.

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

Microsoft Corporation. (2019). Microsoft Excel [Computer software]. Microsoft Corporation.

Ministry of Education, Bangladesh. (2012). National Education Policy 2010. Ministry of Education.

Nunan, D. (1991). Communicative tasks and the language curriculum. TESOL Quarterly, 25(2), 279-295.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). McGraw-Hill.

Pallant, J. (2016). SPSS survival manual: A step by step guide to data analysis using IBM SPSS (6th ed.). Allen & Unwin.

Pastore, S., & Andrade, H. L. (2019). Teacher assessment literacy: A three-dimensional model. Teaching and Teacher Education, 84, 128–138. https://doi.org/10.1016/j.tate.2019.05.003

Patton, K., & Parker, M. (2017). Teacher education communities of practice: More than a culture of collaboration. Teaching and Teacher Education, 67, 351–360. https://doi.org/10.1016/j.tate.2017.06.013

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R.

F. Krueger (Eds.), Handbook of research methods in personality psychology (pp. 224--239). The Guilford Press.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. Educational Measurement: Issues and Practice, 12(4), 10--12, 39. https://doi.org/10.1111/j.1745-3992.1993.tb00548.x

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology, 88(5), 879--903. https://doi.org/10.1037/0021-9010.88.5.879

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

Popham, W. J. (2011). Transformative assessment in action: An inside look at applying the process. ASCD.

QSR International. (2020). NVivo (Version 12) [Computer software]. QSR International Pty Ltd.

Qualtrics. (2020). Qualtrics XM [Online survey platform]. Qualtrics.

R Core Team. (2020). R: A language and environment for statistical computing (Version 4.0) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rahman, M. M., & Pandian, A. (2018). A Critical Investigation of English Language Teaching in Bangladesh: Unfulfilled expectations after two decades of Communicative Language

Teaching. English Today, 34(3), 43–49. doi:10.1017/S026607841700061X

Rahman, M. M., Singh, M. K. M., & Karim, A. (2023). Exam-centric assessment culture in Bangladeshi higher education: Challenges and reform. Higher Education Research & Development, 42(4), 897--912. https://doi.org/10.1080/07294360.2022.2118211

Richards, J. C., & Rodgers, T. S. (2014). Approaches and methods in language teaching (3rd ed.). Cambridge University Press.

Rindfleisch, A., Malter, A. J., Ganesan, S., & Moorman, C. (2008). Cross-sectional versus longitudinal survey research: Concepts, findings, and guidelines. Journal of Marketing Research, 45(3), 261–279. https://doi.org/10.1509/jmkr.45.3.261

Romano, J., Kromrey, J., Coraggio, J. & Skowronek, J. (2006). Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen'sd for evaluating group differences on the NSSE and other surveys?. In annual meeting of the Florida Association of Institutional Research (pp. 1--3).

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. Practical Assessment, Research, and Evaluation, 11(10). https://doi.org/10.7275/9wph-vv65

Sanchez, L., Penarreta, J. & Soria Poma, X. Learning management systems for higher education: a brief comparison. Discov Educ 3, 58 (2024). https://doi.org/10.1007/s44217-024-00143-5

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. Language Testing, 30(3), 309–327.

https://doi.org/10.1177/0265532213480128

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. Annual Review of Sociology, 29, 65--88. https://doi.org/10.1146/annurev.soc.29.110702.110112

Sedgwick, P. (2014). Cross sectional studies: Advantages and disadvantages. BMJ, 348, g2276. https://doi.org/10.1136/bmj.g2276

Shepard, L. A. (2000). The role of assessment in a learning culture. Educational Researcher, 29(7), 4--14. https://doi.org/10.3102/0013189X029007004

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

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 15(2), 4--14. https://doi.org/10.3102/0013189X015002004

standards-based education reform. TESOL Quarterly, 48(3), 586--614.

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

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

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

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and

internal consistency. Journal of Personality Assessment, 80(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Sultana, N. (2019). Language assessment literacy: an uncharted area for the English language teachers in Bangladesh. Lang Test Asia 9, 1. https://doi.org/10.1186/s40468-019-0077-8

Tabachnick, B. G., & Fidell, L. S. (2013). Using multivariate statistics (6th ed.). Pearson.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Taylor, L. (2009). Developing assessment literacy. Annual Review of Applied Linguistics, 29, 21--36. https://doi.org/10.1017/S0267190509090035

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

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. Trends in Sport Sciences, 21(1), 19–25.

Tongco, M. D. C. (2007). Purposive sampling as a tool for informant selection. Ethnobotany Research and Applications, 5, 147–158.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. Cambridge University Press.

Trochim, W. M., & Donnelly, J. P. (2008). The research methods knowledge base (3rd ed.). Cengage Learning.

Tsagari, D. (2021). Language assessment literacy: A decade of research and beyond. Language Testing in Asia, 11, 1–7. https://doi.org/10.1186/s40468-021-00130-5

Tummons, J. (2024). Communities of practice in higher education: Learning, teaching, and research. Routledge. https://doi.org/10.4324/9781003412106

Turnitin, LLC. (2020). Turnitin [Plagiarism detection software]. Turnitin, LLC.

University Grants Commission of Bangladesh. (2015). Quality assurance and accreditation in higher education. UGC.

Van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities. PLoS Medicine, 2(10), e267. https://doi.org/10.1371/journal.pmed.0020267

Vescio, V., Ross, D., & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. Teaching and Teacher Education, 24(1), 80–91. https://doi.org/10.1016/j.tate.2007.01.004

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

Wang, X., & Cheng, Z. (2020). Cross-sectional studies: Strengths, weaknesses, and recommendations. Chest, 158(1S), S65–S71. https://doi.org/10.1016/j.chest.2020.03.012

Wickham, H., & Bryan, J. (2025). readxl: Read Excel files (Version 1.4.5) [R package]. https://readxl.tidyverse.org Readxl

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686 joss.theoj.org

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2025). dplyr: A grammar of data manipulation (Version 1.1.4) [R package]. https://dplyr.tidyverse.org dplyr.tidyverse.org

Wickham, H., Hester, J., & Bryan, J. (2024). readr: Read rectangular text data (Version 2.1.5) [R package]. https://readr.tidyverse.org readr.tidyverse.org

Wickham, H., Vaughan, D., & Girlich, M. (2025). tidyr: Tidy messy data (Version 1.3.1) [R package]. https://tidyr.tidyverse.org tidyr.tidyverse.org

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin, 1(6), 80–83. https://doi.org/10.2307/3001968

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

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

Zhang, Z., & Burry-Stock, J. A. (1997, March 24--28). Assessment practices inventory: A multivariate analysis of teachers' perceived assessment competency [Paper presentation].

Annual Meeting of the National Council on Measurement in Education, Chicago, IL, United States. https://files.eric.ed.gov/fulltext/ED408333.pdf

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. Applied Measurement in Education, 16(4), 323--342. https://doi.org/10.1207/S15324818AME1604_4

Appendix

Appendix 1

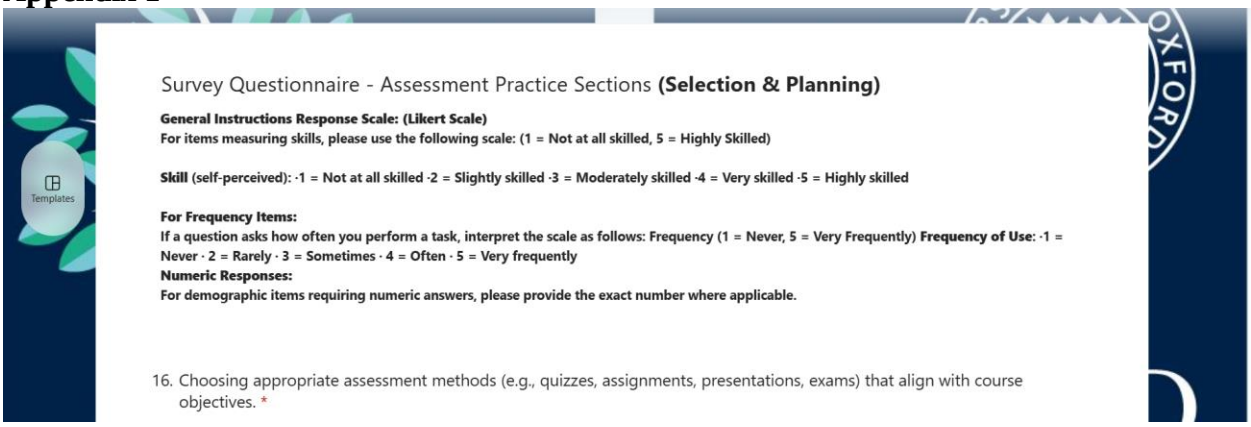


Figure 1. Visual interface in screenshot of the survey questionnaire in Microsoft Forms.

Content of Survey Questionnaire

Adapted for Undergraduate English Teachers at Private and Public Universities in Bangladesh

Dear Participant,

Thank you for taking the time to complete this survey. The purpose of this inventory is to gather detailed insights into your skills and practices in language assessment. Your responses will help inform research and professional development initiatives. Please answer each item honestly. Your responses will remain confidential.

General Instructions

Response Scale: (Likert Scale)

For items measuring skills, please use the following scale: (1 = Not at all skilled, 5 = Highly Skilled)

Skill (self-perceived):

- 1 = Not at all skilled
- 2 = Slightly skilled
- 3 = Moderately skilled
- 4 = Very skilled
- 5 = Highly skilled

For Frequency Items:

If a question asks how often you perform a task, interpret the scale as follows: Frequency (1 = Never, 5 = Very Frequently)

Frequency of Use:

- 1 = Never
- 2 = Rarely
- 3 = Sometimes
- 4 = Often
- 5 = Very frequently

Numeric Responses:

For demographic items requiring numeric answers, please provide the exact number where applicable.

Demographic Information

1. Type of Institution: Public / Private

2. Years of Teaching Experience: ____ years

3. Gender (optional): Male / Female / Other / Prefer not to say

4. Highest Degree Obtained: Bachelor’s / Master’s / M.Phil. / Ph.D. / Other

5. Academic Position/Rank: Lecturer / Senior Lecturer / Assistant Professor / Associate Professor / Professor / Other

6. Region (Workplace Location): Dhaka / Chittagong / Rajshahi / Khulna / Sylhet / Barisal / Rangpur / Mymensingh / Other

7. Formal Training in Assessment: Yes / No

Assessment Practice Sections

Selection & Planning

1. Choosing appropriate assessment methods (e.g., quizzes, assignments, presentations, exams) that align with course objectives.

Frequency (1–5): _____ Skill (1–5): _____

2. Administering scheduled quizzes (quizzes announced in advance).

Frequency (1–5): _____ Skill (1–5): _____

3. Administering surprise quizzes (without prior announcement).

Frequency (1–5): _____ Skill (1–5): _____

4. Assessing student understanding based on their in-class questions.

Frequency (1–5): _____ Skill (1–5): _____

5. Observing students during class to assess their learning or participation.

Frequency (1–5): _____ Skill (1–5): _____

6. Using a test blueprint (table of specifications) to plan assessments.

Frequency (1–5): _____ Skill (1–5): _____

7. Designing assessments based on clearly defined course objectives.

Frequency (1–5): _____ Skill (1–5): _____

8. Ensuring that assessments align with instruction.

Frequency (1–5): _____ Skill (1–5): _____

9. Ensuring exams sample the full range of course content.

Frequency (1–5): _____ Skill (1–5): _____

Construction of Traditional Tests

10. Selecting or adapting textbook-provided test items.

Frequency (1–5): _____ Skill (1–5): _____

11. Revising previously used tests to suit current instruction.

Frequency (1–5): _____ Skill (1–5): _____

12. Designing written tests such as midterms or finals.

Frequency (1–5): _____ Skill (1–5): _____

13. Writing multiple-choice questions.

Frequency (1–5): _____ Skill (1–5): _____

14. Writing matching questions.

Frequency (1–5): _____ Skill (1–5): _____

15. Writing true/false questions.

Frequency (1–5): _____ Skill (1–5): _____

16. Writing fill-in-the-blank or short answer questions.

Frequency (1–5): _____ Skill (1–5): _____

17. Writing essay questions.

Frequency (1–5): _____ Skill (1–5): _____

18. Creating questions for higher-order thinking.

Frequency (1–5): _____ Skill (1–5): _____

19. Developing model answers or rubrics for essay scoring.

Frequency (1–5): _____ Skill (1–5): _____

Performance-Based Assessment

20. Aligning performance tasks with instruction and objectives.

Frequency (1–5): _____ Skill (1–5): _____

21. Defining performance criteria in advance using a rubric.

Frequency (1–5): _____ Skill (1–5): _____

22. Sharing performance assessment criteria with students before task.

Frequency (1–5): _____ Skill (1–5): _____

23. Using a checklist/rating scale during performance observations.

Frequency (1–5): _____ Skill (1–5): _____

24. Using concept maps as an assessment method.

Frequency (1–5): _____ Skill (1–5): _____

25. Assessing individual class participation.

Frequency (1–5): _____ Skill (1–5): _____

26. Assessing group participation.

Frequency (1–5): _____ Skill (1–5): _____

27. Assessing individual hands-on assignments or tasks.

Frequency (1–5): _____ Skill (1–5): _____

28. Assessing group projects or collaborative work.

Frequency (1–5): _____ Skill (1–5): _____

29. Evaluating student presentations.

Frequency (1–5): _____ Skill (1–5): _____

30. Using portfolios to assess student progress.

Frequency (1–5): _____ Skill (1–5): _____

Interpretation of Standardized Tests

31. Determining validity of standardized tests for your students.

Frequency (1–5): _____ Skill (1–5): _____

32. Following procedures when administering standardized tests.

Frequency (1–5): _____ Skill (1–5): _____

33. Interpreting test scores like percentiles to students and parents.

Frequency (1–5): _____ Skill (1–5): _____

34. Explaining standardized English test scores (e.g., IELTS/TOEFL).

Frequency (1–5): _____ Skill (1–5): _____

35. Using test scores to diagnose learning needs.

Frequency (1–5): _____ Skill (1–5): _____

Use of Assessment Results

36. Calculating and interpreting test statistics (mean, SD, etc.).

Frequency (1–5): _____ Skill (1–5): _____

37. Conducting item analysis for classroom tests.

Frequency (1–5): _____ Skill (1–5): _____

38. Revising tests based on item analysis.

Frequency (1–5): _____ Skill (1–5): _____

39. Using assessment results to guide teaching.

Frequency (1–5): _____ Skill (1–5): _____

40. Using results to improve syllabus or curriculum.

Frequency (1–5): _____ Skill (1–5): _____

41. Using results for student placement or promotion.

Frequency (1–5): _____ Skill (1–5): _____

42. Evaluating class performance improvement using assessments.

Frequency (1–5): _____ Skill (1–5): _____

43. Evaluating department or program improvement using assessments.

Frequency (1–5): _____ Skill (1–5): _____

Grading Practices

44. Creating systematic grading procedures.

Frequency (1–5): _____ Skill (1–5): _____

45. Developing a personal grading philosophy.

Frequency (1–5): _____ Skill (1–5): _____

46. Using norm-referenced grading (grading on a curve).

Frequency (1–5): _____ Skill (1–5): _____

47. Using criterion-referenced grading (grading against a standard).

Frequency (1–5): _____ Skill (1–5): _____

48. Having procedures for borderline grades.

Frequency (1–5): _____ Skill (1–5): _____

49. Informing students in advance about grading policies.

Frequency (1–5): _____ Skill (1–5): _____

50. Establishing grading expectations for special needs students.

Frequency (1–5): _____ Skill (1–5): _____

51. Weighting exams, assignments, and participation differently.

Frequency (1–5): _____ Skill (1–5): _____

52. Including extra credit activities in grading.

Frequency (1–5): _____ Skill (1–5): _____

53. Considering student ability in grading.

Frequency (1–5): _____ Skill (1–5): _____

54. Considering classroom behavior in grading.

Frequency (1–5): _____ Skill (1–5): _____

55. Considering improvement in grading.

Frequency (1–5): _____ Skill (1–5): _____

56. Considering effort in grading.

Frequency (1–5): _____ Skill (1–5): _____

57. Considering attendance in grading.

Frequency (1–5): _____ Skill (1–5): _____

58. Assigning final grades.

Frequency (1–5): _____ Skill (1–5): _____

Communication and Ethics

59. Providing oral feedback to students.

Frequency (1–5): _____ Skill (1–5): _____

60. Providing written feedback on student work.

Frequency (1–5): _____ Skill (1–5): _____

61. Communicating assessment results to students.

Frequency (1–5): _____ Skill (1–5): _____

62. Communicating assessment results to parents or guardians.

Frequency (1–5): _____ Skill (1–5): _____

63. Communicating assessment results to fellow educators.

Frequency (1–5): _____ Skill (1–5): _____

64. Avoiding ‘teaching to the test’.

Frequency (1–5): _____ Skill (1–5): _____

65. Protecting student confidentiality in test results.

Frequency (1–5): _____ Skill (1–5): _____

66. Identifying unethical assessment practices.

Frequency (1–5): _____ Skill (1–5): _____

67. Identifying unethical uses of assessment information.

Frequency (1–5): _____ Skill (1–5): _____

Final Open-Ended Questions

1. Challenges: "What challenges do you face in assessing students effectively?"
2. Training Needs: "What kind of assessment training or professional development would benefit you the most?"

Appendix 2

Table 10. Wilcoxon Signed-Rank Test Results for Paired Items of Perceived Skill and Reported Use (N = 52)

Item Text	n	Mdn Skill	Mdn Use	W	p	p_FDR	r
Selection & Planning							
1. Choosing appropriate assessment methods	18	4	4	58.0	0.198	0.233	.166
2. Administering scheduled quizzes	14	4	4	24.0	0.058	0.086	.248
3. Administering surprise quizzes	23	3	2	28.0	<.001	0.005	.463*
4. Observing students during instruction	17	4	4	30.0	0.018	0.034	.305*
5. Using a test blueprint (table of specifications)	19	4	4	90.5	0.843	0.843	.025
6. Ensuring assessments cover course content	16	3	3	29.0	0.031	0.054	.279
7. Designing assessments based on course objectives	10	4	4	25.0	0.782	0.793	.035
8. Ensuring that assessments align with instruction	9	4	4	9.0	0.083	0.116	.221

9. Using various assessment formats	11	4	4	15.0	0.088	0.118	.221
Construction of Traditional Tests							
10. Selecting/adapting items from textbooks	17	4	4	44.0	0.111	0.146	.213
11. Revising old tests	15	4	4	11.0	0.004	0.011	.385*
12. Designing midterm/final exams	7	5	5	9.0	0.380	0.410	.117
13. Writing multiple-choice questions	26	4	3	3.5	<.001	<.001	.511*
14. Writing matching questions	26	4	3	15.0	<.001	0.005	.441*
15. Writing true/false questions	24	4	3	10.0	<.001	0.002	.483*
16. Writing fill-in-the-blank questions	25	4	4	45.0	0.005	0.014	.392*
17. Writing short-answer essay questions	17	4	4	51.0	0.334	0.369	.128
18. Writing higher-order thinking questions	23	4	4	66.0	0.077	0.108	.244
19. Developing model answers for essay scoring	19	4	4	76.0	0.388	0.418	.114

Performance-Based Assessment							
20. Aligning performance tasks with instruction	14	4	4	27.5	0.180	0.216	.178
21. Defining criteria for performance with rubrics	21	4	4	66.0	0.080	0.111	.238
22. Sharing performance criteria with students	18	4	4	45.0	0.041	0.070	.273
23. Using checklists and/or rating scales	23	4	4	46.0	0.009	0.021	.357*
24. Using concept maps for assessment	19	3	2	10.0	<.001	0.002	.464*
25. Assessing individual participation	18	4	4	34.5	0.014	0.028	.324*
26. Assessing hands-on assignments	20	4	4	35.0	0.004	0.011	.399*
27. Assessing group projects	16	4	4	44.0	0.264	0.296	.148
28. Assessing oral presentations	20	4	4	19.0	<.001	0.004	.445*
29. Using portfolios to assess student learning	22	4	3	15.0	<.001	<.001	.492*

30. Creating authentic performance tasks	17	4	3	19.0	<.001	0.004	.445*
Interpretation of Standardised Tests							
31. Determining validity of standardised tests	20	4	4	19.0	<.001	0.004	.445*
32. Interpreting percentiles to students/parents	26	4	4	19.0	<.001	0.004	.445*
33. Explaining IELTS/TOEFL scores	24	4	3	5.0	<.001	0.005	.469*
34. Using test scores to diagnose learning needs	26	4	3	15.0	<.001	0.005	.466*
35. Calculating and interpreting test statistics	20	3	3	15.0	<.001	0.005	.465*
Use of Assessment Results							
36. Conducting item analysis	22	3	2	0.0	<.001	0.004	.530*
37. Revising tests based on item analysis	21	3	3	46.0	0.034	0.058	.289
38. Using assessment results to guide teaching	22	4	3	21.0	<.001	0.002	.464*

39. Using results to improve syllabus/curriculum	19	4	3	15.0	<.001	0.005	.465*
40. Using results for student placement/promotion	21	4	3	6.0	<.001	0.005	.445*
41. Evaluating class performance using assessments	18	4	4	21.0	0.001	0.005	.434*
42. Reporting assessment results to stakeholders	19	4	3	15.0	<.001	0.005	.465*
43. Using data to inform institutional decisions	18	3	3	36.0	0.041	0.070	.273
Grading & Feedback Practices							
44. Creating systematic grading procedures	15	4	4	30.0	0.082	0.114	.228
45. Developing a personal grading philosophy	14	4	4	21.0	0.049	0.078	.263
46. Using norm-referenced grading (on a curve)	19	3	2	10.0	<.001	0.002	.464*
47. Using criterion-referenced grading	17	4	4	17.0	0.002	0.007	.410*
48. Having procedures for borderline grades	18	3	3	27.0	0.004	0.011	.385*

49. Informing students about grading policies	9	4	4	9.0	0.083	0.116	.221
50. Establishing grading for special needs students	16	3	3	20.0	0.009	0.021	.344*
51. Weighting exams, assignments, participation	11	4	4	15.0	0.088	0.118	.221
52. Including extra credit in grading	20	3	3	35.0	0.004	0.011	.399*
53. Considering student ability in grading	19	4	3	15.0	<.001	0.005	.465*
54. Considering classroom behaviour in grading	26	3	3	4.0	<.001	0.004	.492*
55. Considering improvement in grading	24	4	3	10.0	<.001	0.002	.483*
56. Considering effort in grading	22	4	4	36.0	0.013	0.027	.328*
57. Considering attendance in grading	21	4	4	36.0	0.008	0.019	.365*
58. Assigning final grades	8	5	5	10.0	0.164	0.203	.174
Communication & Ethics							
59. Providing oral feedback	18	4	4	27.0	0.004	0.011	.385*

60. Providing written feedback	17	4	4	25.5	0.008	0.019	.352*
61. Communicating results to students	15	4	4	30.0	0.082	0.114	.228
62. Communicating results to parents/guardians	26	3	2	0.0	<.001	0.004	.530*
63. Communicating results to fellow educators	25	3	3	25.0	<.001	0.001	.506*
64. Communicating results to administrators	20	3	3	30.0	0.001	0.005	.434*
65. Communicating results to external stakeholders	18	3	2	15.0	<.001	0.002	.464*
66. Avoiding 'teaching to the test'	23	4	4	46.0	0.009	0.021	.357*
67. Protecting student confidentiality	10	5	5	15.0	0.500	0.526	.080

Note. The table presents results for all 67 assessment literacy items. *n* = number of pairs with non-zero differences. *Mdn* = Median. *W* = Wilcoxon test statistic. *p* = uncorrected *p*-value. *p*_FDR = *p*-value adjusted for false discovery rate using the Benjamini-Hochberg procedure. *r* = effect size. Data derived from survey responses of 52 Bangladeshi undergraduate English teachers.

*p*_FDR < .05.

Appendix 3

Exploring Assessment Literacy of Undergraduate English Language Teachers in Bangladesh: Practices and Challenges

PARTICIPANT INFORMATION SHEET

Central University Research Ethics Committee Reference: [Education (Educ) DREC - 1891578]

Introductory paragraph

You are being invited to take part in a research project. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information. Take time to decide whether you wish to take part.

Why is this research being conducted?

This study investigates the assessment literacy of undergraduate English language teachers working in Bangladeshi universities. It aims to explore teachers' self-reported assessment skills, practices, and challenges, and to identify areas where professional development is needed to improve assessment quality in higher education.

Why have I been invited to take part?

You have been invited because you are an English language teacher currently teaching at the undergraduate level in a Bangladeshi public or private university. The study seeks input from approximately 50–100 participants who meet this criterion.

Do I have to take part?

No. It is entirely your decision whether or not to participate. You may withdraw from the study at any time before submitting the survey, without giving a reason and without any consequences. Once your response is submitted anonymously, it cannot be withdrawn, as it will not be possible to identify your data.

What will happen to me if I take part in the research?

You will be asked to complete an online survey via the Qualtrics platform. The survey includes multiple-choice, Likert-scale, and open-ended questions about your assessment practices and experiences. It will take approximately 20–25 minutes to complete. You may skip any question or exit the survey at any time before submission. No interviews, recordings, or follow-up sessions are involved.

What are the possible disadvantages and risks in taking part?

There are no foreseeable risks or disadvantages to participating. The survey is anonymous and does not ask for personal or sensitive information. Your responses will not be identifiable and will be used solely

for academic research.

Are there any benefits in taking part?

There is no direct benefit to you personally. However, your contribution will help inform improvements in assessment-related training and practices for language teachers in higher education in Bangladesh.

Expenses and payments

There will be no payment for taking part in this research.

What information will be collected and why is the collection of this information relevant for achieving the research objectives?

The survey will collect information on your teaching experience, assessment skills, and use of different assessment methods. This includes Likert-scale ratings and open-ended responses. The information will help identify current practices, gaps between knowledge and implementation, and training needs. Data will be stored securely on University of Oxford systems for a maximum of five years. The researcher and supervisor will have access to anonymised data. Consent forms (for paper-based surveys only) will be stored separately in a locked cabinet accessible only to the researcher.

Will the research be published? Could I be identified from any publications or other research outputs?

The results of the research may be presented in a master's dissertation, academic publications, and conference presentations. All data will be anonymised and no participant will be identified. No direct quotations will be used from the open-ended responses.

Data Protection

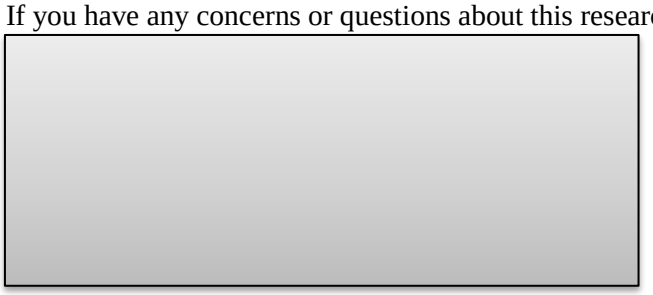
The University of Oxford is the data controller with respect to your personal data, and as such will determine how your personal data is used in the research. The University will process your personal data for the purpose of the research outlined above. Research is a task that is performed in the public interest. Further information about your rights with respect to your personal data is available from the University's Information Compliance website at https://compliance.admin.ox.ac.uk/individual-rights.

Who has reviewed this research?

This research has received favourable opinion from a subcommittee of the University of Oxford Central University Research Ethics Committee. (Ethics reference: **xxxxx)**.

Who do I contact if I have a concern about the research or I wish to complain?

If you have any concerns or questions about this research, please contact:



Email: md.yasir@education.ox.ac.uk

Supervisor:
Dr Michelle Meadows
Email: michelle.meadows@education.ox.ac.uk

If you remain unhappy or wish to make a formal complaint, please contact:
University of Oxford Research Governance, Ethics & Assurance (RGEA) team
Email: rgea.complaints@admin.ox.ac.uk
Tel: +44 (0)1865 616480
.

Further Information and Contact Details

Appendix 4

Consent to take part in Exploring Assessment Literacy of Undergraduate English Language Teachers in Bangladesh: Practices and Challenges

Central University Research Ethics Committee (CUREC) reference: Education (Educ) DREC - 1891578

Purpose of Study: This study aims to explore the assessment practices, perceived skills, challenges, and professional development needs of undergraduate English language teachers in Bangladeshi universities, using an anonymous survey to collect both quantitative and qualitative data.

	Please initial each box if you agree with the statement
I confirm that I have read and understood the Participant Information Sheet (version 5.7, dated June 2025) for the above research. I have had the opportunity to consider the information, ask questions, and have had these answered satisfactorily.	
I understand that my participation is voluntary and that I am free to withdraw at any point before submitting the survey, without giving any reason and without any consequence.	
I understand that the survey is anonymous and that no identifying information will be collected or stored.	
I understand who will have access to personal data provided, how the data will be stored and what will happen to the data at the end of the project.	
I understand that I will **not** be identifiable from any publications or other research outputs, including academic reports, conference presentations, or the researcher's dissertation. All responses will be collected anonymously and presented in a way that ensures individuals cannot be identified.	
I understand that I will not receive individual feedback but may request a summary of overall findings through a separate, non-linked process.	
I understand how to raise a concern or make a complaint.	
I agree to take part in this study.	

_____ dd / mm / yyyy _____

Name of participant	Date	Signature
_____	dd / mm / yyyy	_____
Name of person taking consent	Date[1]	Signature

[1] To be signed and dated in the presence of the participant. Once this has been signed by both parties the participant should receive a copy of the signed and dated participant consent form. The original signed and dated consent form should be kept with the project's main documents, which must be kept in a secure location.

Appendix 5 (next page)

Ethics reference: Education (Educ) DREC – 1891578

Education (Educ) DREC
15 Norham Gardens, Oxford, OX2 6PY

Applicant: [redacted]
Principal Investigator: [redacted]
Department: Education

Study title: Exploring Assessment Literacy of Undergraduate English Language Teachers in Bangladesh: Practices and Challenges
(Version: 1.0)
Ethics reference: Education (Educ) DREC - 1891578

Dear [redacted]
On behalf of the Committee, I confirm that the above research study described in the application and other supporting documentation submitted to the committee has been carefully considered by the Education (Educ) DREC in accordance with the University's regulations and policy for ethics approval of research involving human participants, human tissue and/or personal data. The opinion is as follows:

Opinion of Research Ethics Committee: Favourable Opinion
Subject to the following conditions:
Decision Date: 14 Jul 2025, 09:59
Opinion End Date: 13 Jan 2027

If favourable, insurance-provided indemnity arrangements will be in place between the decision date and opinion end date and you may now commence your study activities. Should you plan to continue the research beyond the end date above, it is your responsibility to ensure that you request, and receive, an extension (via amendment) from the committee for indemnity to remain in place. You may be required to provide a justification.
Please note the following:
Amendments: Should there be any subsequent changes to the reviewed study, applications for amendments can be made via the Oxford Ethics Application System (Worktribe Ethics).
Reports: Studies considered by OxTREC are expected to submit an *annual progress report* on each anniversary of study approval, until the study is completed. An end of study report is also required.
Audit: This study may be selected for audit at the discretion of the Research Governance, Ethics and Assurance Team.
Data safety: It is the responsibility of the PI to ensure that all data collected during the course of the study is stored and transferred safely and securely in accordance with University requirements. Further guidance and advice are available from the Research Data Team. Additional information is available at https://researchsupport.web.ox.ac.uk/governance/ethics
Yours Sincerely

Education Ethics Officer