

Development and Standardization of Testing Systems for English Listening and Speaking: A Framework for Bangladeshi Higher Education

Md. Niaz Alamgir

Assistant Professor of English,

Institute of Modern Languages (IML), Jagannath University,
Dhaka, Bangladesh

E-mail: niazalamgir@iml.jnu.ac.bd

Abstract

Listening and speaking are central to communicative competence but remain marginal in many tertiary English assessments in Bangladesh, which continue to privilege written skills and discrete linguistic knowledge. This qualitative documentary study combines a scoping review of empirical research with analysis of policy and test documents to identify gaps in oral-skills assessment, synthesize validity threats, and develop a context-responsive framework for standardization. The synthesis identifies four interrelated constraints: weakly shared proficiency standards, limited assessment literacy and rater preparation, narrow task and scoring practices, and resource and governance limitations. Drawing on contemporary validity theory, the Standards for Educational and Psychological Testing, and design principles from IELTS, TOEFL iBT, and PTE Academic, the paper proposes a six-stage pathway: construct definition; test specifications; task and rubric design; piloting and analysis; standard setting and reporting; and implementation with continuing quality assurance. The framework emphasizes national coordination, transparent validity arguments, localized communicative demands, rater calibration, moderation, audit trails, and scalable technology-supported administration. It offers a practical basis for improving the fairness, reliability, comparability, and educational value of listening and speaking assessment in Bangladeshi higher education.

Keywords: Bangladesh; Listening assessment; Speaking assessment; Standardization; Validity

1. Introduction

English occupies a central place in contemporary higher education, shaping access to disciplinary knowledge, research communication, and graduate employability. In Bangladesh, the expansion of private universities, the growth of English-medium and bilingual programs, and the global orientation of the labor market have intensified expectations that university graduates can participate in academic and professional interactions in English (World Bank, 2019). Despite these demands, assessment practices across many institutions still mirror a long-standing pattern in the wider education system: examinations emphasize reading, writing, and decontextualized grammatical knowledge, while listening and speaking receive limited or inconsistent attention (Al Amin & Greenwood, 2018; Islam et al., 2021).

Assessment is not merely a measurement tool; it also drives teaching and learning. Consequently, under-assessment of oral skills can reproduce under-instruction. When listening and speaking are excluded from high-stakes decisions, classroom time tends to prioritize what is tested, narrowing opportunities for interaction, reducing the perceived value of communicative pedagogy, and weakening students' confidence in real-world communication (Alderson & Wall, 1993; Cheng, 2005). For many Bangladeshi students, tertiary education is the final institutional stage before entry into workplaces where spoken English is increasingly expected. Deficiencies in oral-skills assessment are therefore not only an academic concern but also a social and economic issue connected to fairness and opportunity (McNamara & Roever, 2006).

This paper responds to a practical and scholarly problem: how can tertiary institutions in Bangladesh develop and standardize assessment systems for listening and speaking that are defensible, feasible, and locally meaningful? Rather than proposing a single national test as a quick solution, the paper treats standardization as a set of shared practices: common construct definitions, transparent test specifications, calibrated scoring procedures, and comparable reporting that allow scores to be interpreted consistently across contexts (American Educational Research Association et al., 2014; Bachman & Palmer, 2010).

The study addresses three research questions:

- Q1. What challenges hinder standardized assessment of listening and speaking skills at the tertiary level in Bangladesh?
- Q2. How do prevailing assessment practices compare with international testing models and quality assurance procedures?
- Q3. What framework can support reliable, valid, and contextually grounded oral assessment in Bangladeshi higher education?

2. Background and Problem Context

Bangladesh's assessment culture has often been described as examination-oriented, with high-stakes testing shaping instructional priorities and learners' study strategies. At the macro level, this pattern has been linked to curricular crowding, limited formative assessment practices, and a tendency to treat assessment primarily as a summative gatekeeping mechanism (Al Amin & Greenwood, 2018; World Bank, 2019). In language education, a further tension emerges between policy aspirations for communicative language teaching (CLT) and the realities of classroom practice, especially in resource-constrained environments. Policy-level shifts toward CLT have not consistently translated into oral-skills assessment systems that can support communicative outcomes (Hamid & Baldauf, 2008; Rahman & Pandian, 2018).

The school-to-university transition is particularly important because many students arrive at tertiary institutions with limited exposure to extended English interaction. A large baseline study of listening and speaking in Bangladesh's higher secondary schools reported widespread concern about oral proficiency and documented contextual constraints such as limited practice opportunities and fear of making mistakes (Brunfaut & Green, 2017). These constraints are not automatically resolved at the tertiary level. Universities vary widely in class size, teacher profiles, infrastructure, and assessment policies. In some contexts, oral assessment is conducted informally or on an ad hoc basis, whereas in others it is included but lacks common rubrics, rater training, or moderation procedures that would support comparability across cohorts (Islam et al., 2021).

In such circumstances, the central validity concern is not simply whether a speaking or listening task appears “communicative” on the surface, but whether the resulting scores can be interpreted reliably and fairly for their intended uses. Variability in tasks, scoring criteria, and administration conditions may introduce construct-irrelevant variance, while the absence of systematic task sampling may under-represent the construct of oral proficiency (Messick, 1989). These issues are amplified in high-stakes contexts in which oral assessments inform course completion, progression, or graduation decisions.

Standardization, therefore, should be framed as an institutional and national capacity-building agenda. It includes building shared proficiency standards, developing assessment literacy, training raters and examiners, and establishing quality assurance mechanisms appropriate to local constraints. International models can offer design exemplars, but they must be adapted to Bangladeshi curricular goals, sociolinguistic realities, and practical conditions (Bachman & Palmer, 2010; Fulcher & Davidson, 2007).

3. Theoretical and Conceptual Framework

The conceptual foundation of this paper draws on three interrelated traditions: communicative competence and performance assessment; contemporary validity theory; and washback and the social consequences of testing. Together, these traditions support a view of listening and speaking assessment as an evidence-based interpretive practice rather than a purely technical procedure.

First, communicative language testing understands proficiency as the ability to use language appropriately and effectively in context. Classic models of communicative competence emphasize not only grammatical knowledge but also sociolinguistic, discourse, and strategic abilities (Canale & Swain, 1980). Subsequent assessment frameworks conceptualize language ability as an interaction between language knowledge, strategic competence, and contextual features of tasks (Bachman, 1990; Bachman & Palmer, 1996). For speaking assessment, this implies that elicitation must capture both individual production and interactional processes, such as turn-taking, alignment, and repair. Recent work on interactional competence highlights that speaking

performance is co-constructed with interlocutors and shaped by test format, which has direct implications for construct definition and rating scale design (Galaczi & Taylor, 2018; Luoma, 2004).

For listening assessment, theoretical accounts stress that comprehension is an active meaning-making process involving linguistic decoding, discourse processing, pragmatic inference, and the integration of prior knowledge. Listening tasks must therefore be designed with careful attention to text type, speaker characteristics, and task demands (Buck, 2001). If listening is assessed only through short, decontextualized items, the construct may be reduced to recognition of discrete information rather than comprehension in authentic academic or workplace contexts.

Second, validity is conceptualized as an argument about the meaning and use of test scores. In Messick's (1989) unified view, validity concerns both evidential and consequential aspects of assessment, including its social impact. Kane (2013) further articulates validity as an interpretive argument that links observed performances to score-based decisions through warrants, assumptions, and evidence. This orientation is reflected in the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014), which require evidence to support intended score interpretations and uses. For oral skills, a validity argument typically draws on evidence from content representation, response processes, internal structure (e.g., rater consistency), relationships to external criteria, and consequences (Chapelle et al., 2008).

Third, washback research demonstrates that tests can shape curricula, teaching practices, and learner motivation, sometimes in unintended ways. Alderson and Wall's (1993) washback hypotheses and subsequent empirical work suggest that high-stakes tests may narrow curricula unless assessment systems intentionally align tasks, criteria, and feedback practices with desired learning outcomes (Cheng, 2005; Green, 2014). In Bangladesh, where examinations have historically prioritized written skills, a shift toward integrated oral assessment has the potential to support communicative pedagogy. This potential can be realized only when teachers and students perceive assessment practices as legitimate, transparent, and fair (Kunnan, 2004; McNamara & Roever, 2006).

4. Literature Review

The literature relevant to oral-skills assessment in Bangladesh spans three domains: (a) studies documenting assessment practices and challenges in Bangladesh, (b) research on speaking and listening test development and scoring, and (c) scholarship on standardization and quality assurance in performance assessment. Although direct empirical studies of tertiary-level listening and speaking assessment in Bangladesh are limited, existing work and adjacent evidence provide a coherent picture of systemic constraints and development needs.

Assessment practices and constraints in Bangladesh. Available research suggests that language assessment in Bangladesh remains dominated by traditional written examinations, with limited classroom-based formative assessment and insufficient teacher preparation for assessment design and interpretation (Hasan, 2020; Islam et al., 2021). In the broader education sector, accountability pressures and examination traditions have reinforced a “teaching to the test” ecology that can undermine communicative instruction (Al Amin & Greenwood, 2018). Policy and planning scholarship has similarly argued that CLT implementation in Bangladesh is “bogged down” when curricular expectations for communication are not supported by assessment systems, teacher development, and institutional resources (Hamid & Baldauf, 2008; Hamid & Erling, 2016). At the tertiary level, anecdotal and small-scale studies indicate variation across universities in whether and how oral skills are assessed and in the extent to which rating practices are guided by explicit rubrics (Kabir, 2015).

Listening and speaking as assessment constructs. International language assessment research offers detailed guidance on defining and sampling oral constructs. For speaking, important design choices include test format (individual, paired, or group), interlocutor roles, topic selection, and the balance between planned and spontaneous speech. Each choice affects candidates’ opportunities to display their abilities and introduces potential sources of variance that must be addressed in validation (Fulcher, 2015; Luoma, 2004). For listening, task design must consider text authenticity, the degree of interaction (e.g., academic discussions versus monologues), and whether tasks measure comprehension, inferencing, or note-taking skills

(Buck, 2001). In academic contexts, integrated listening-speaking tasks can enhance authenticity by approximating university communication demands (Chapelle et al., 2008).

Standardization and quality assurance. Standardization in performance assessment depends on two related components: transparent specifications and evidence-based scoring. Test specifications define what tasks should look like, what they target, and how they should be administered (Weir, 2005). Scoring standardization requires clear rating scales, rater training, and monitoring systems capable of detecting drift, bias, or inconsistent severity (Fulcher & Davidson, 2007). Many-facet Rasch measurement is frequently used to examine rater effects and strengthen fairness (Fan & Knoch, 2019). Standard setting is also central to interpretation: when tests support proficiency claims such as B1 or B2, developers must establish defensible cut scores through methods such as Angoff, Bookmark, or performance-based standard setting linked to descriptors (Brown & Abeywickrama, 2019; Council of Europe, 2001).

Taken together, the literature suggests that Bangladesh's tertiary-level oral assessment challenges are not unique, but they are intensified by system-level constraints and uneven institutional capacity. The next sections describe how the present documentary and scoping analysis synthesizes these issues and translates them into a staged development framework designed for local adaptation.

5. Method

5.1 Design

The study combined qualitative documentary analysis with a scoping review of relevant evidence. Documentary analysis examined how oral-skills assessment is represented in available reports, studies, and test documentation, while the scoping review mapped the breadth of evidence and identified gaps requiring further empirical research. This approach is appropriate when evidence is heterogeneous and the aim is to clarify concepts and development needs rather than estimate effect sizes (Arksey & O'Malley, 2005; Levac et al., 2010; Tricco et al., 2018).

5.2 Sources of evidence and sampling

Sources included peer-reviewed articles on language assessment in Bangladesh, sector reports on education and assessment, and public documentation on international tests that include listening and speaking (e.g., IELTS, TOEFL iBT, and PTE Academic). While the primary analytic focus was tertiary education, evidence from upper secondary contexts was included when it illuminated pipeline constraints relevant to university assessment (Brunfaut & Green, 2017).

5.3 Data extraction and analysis

Documents were charted for (a) assessment purposes and contexts, (b) definitions of listening and speaking constructs, (c) task types and scoring approaches, (d) reported validity or fairness concerns, and (e) recommendations for improvement. Analysis followed an iterative coding process: open coding to identify recurring challenges and proposed solutions, axial coding to cluster codes into domains (construct definition, tasks, rubrics, administration, rater training, and governance), and interpretive synthesis to develop a staged framework for test development and standardization (Braun & Clarke, 2006).

5.4 Trustworthiness and triangulation

Trustworthiness was addressed through transparent analytic procedures, reflexive memoing, and triangulation. Data-source triangulation combined empirical research, sector and policy reports, and test-provider documentation. Theory triangulation drew on validity theory, communicative competence frameworks, and washback scholarship to interpret recurring patterns (American Educational Research Association et al., 2014; Kane, 2013). Future empirical iterations should add analyst triangulation through double-coding of a document subset and a fuller audit trail of coding decisions (Lincoln & Guba, 1985).

6. Results and Synthesis

The analysis produced four clusters of challenges that repeatedly threaten the validity and reliability of tertiary-level listening and speaking assessment in Bangladesh: construct ambiguity, task and delivery limitations, scoring variability, and governance and capacity constraints. The clusters are

interdependent; improvements in one domain (e.g., better rubrics) are unlikely to yield lasting benefits without parallel improvements in others (e.g., rater training and monitoring).

6.1 A persistent oral-skills gap and its consequences

Across sources, listening and speaking are consistently described as underemphasized in high-stakes assessment, contributing to limited instructional time for oral practice and low learner confidence. In Brunfaut and Green's (2017) baseline study, focus groups identified lack of practice and exposure to English (73% of groups) and shyness or fear of criticism (68%) as major barriers to speaking development. Although the study focused on higher secondary education, these factors may continue into tertiary settings, particularly when universities do not institutionalize oral assessment and supportive feedback practices.

6.2 Construct ambiguity and lack of shared standards

A central standardization problem is the absence of commonly adopted proficiency descriptors for listening and speaking across institutions. When constructs are weakly defined, tasks may drift toward what is easiest to administer rather than what represents target language use. Construct ambiguity also affects fairness: students may be evaluated against implicit norms, such as accent or "native-like" fluency, rather than construct-relevant features aligned with curricular goals (Jenkins, 2015; Kunnan, 2004). Aligning oral constructs with externally recognized descriptors such as CEFR bands can improve comparability, but the alignment must be localized to Bangladeshi communicative demands and instructional realities (Council of Europe, 2001).

6.3 Task design and administration constraints

Documentary evidence indicates that speaking tests, where they exist, often rely on short rehearsed presentations or teacher-led interviews that do not consistently elicit interactional competence. Where listening tests are administered, limited equipment, inconsistent audio quality, and time pressure may affect delivery. These conditions can introduce construct-irrelevant variance, particularly for lower-proficiency students and those with limited exposure to test formats. Because task sampling is a major source of score

generalizability in performance assessment, overreliance on a single task type can under-represent the construct (Bachman & Palmer, 2010; Weir, 2005).

6.4 Scoring variability and rater effects

Speaking assessment is particularly vulnerable to rater-related variance. Without common rating scales and systematic calibration, raters may apply different severity standards or prioritize different features of performance (Fulcher & Davidson, 2007). Analytic rubrics with explicit descriptors can support transparency and instructional feedback, but only when raters develop shared interpretations of scale levels through training, benchmarking, and moderation. Many-facet Rasch measurement can diagnose rater severity, bias, and scale functioning and thereby support fairness-oriented quality assurance (Fan & Knoch, 2019).

6.5 Governance, assessment literacy, and resource capacity

Finally, standardization depends on institutional governance and the professional capacity to design, administer, and interpret assessments. Limited language assessment literacy among teachers and administrators can result in mismatches between stated learning outcomes and assessment practices (Islam et al., 2021). Resource limitations, including large class sizes, limited examiner time, and insufficient recording or computer facilities, further constrain implementation. These constraints are not merely logistical; they shape which constructs can be feasibly assessed and what kinds of evidence can be collected for validation.

Table 1

Reported barriers to students’ English speaking development (community-member focus groups)

Barrier	Focus groups (n)	Percent
Lack of practice/exposure to English	16	73%

Barrier	Focus groups (n)	Percent
Shyness/fear of criticism or mistakes	15	68%
Limited vocabulary	12	55%
Lack of an English-language/cultural environment	11	50%
Grammar weaknesses	9	41%
Pronunciation difficulties	5	23%

Note. Data are reported as presented in Brunfaut and Green (2017, p. 70).

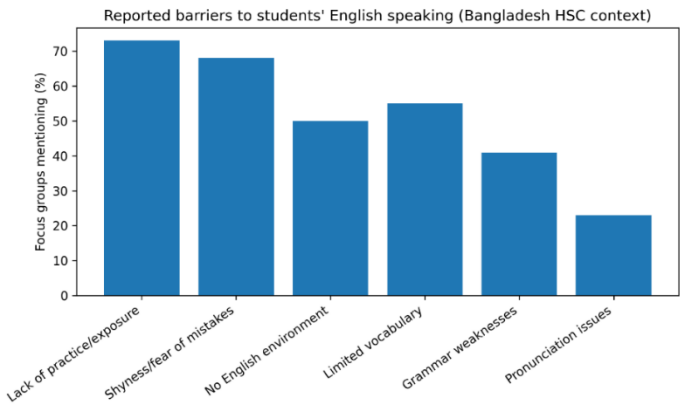


Figure 1
Barriers to students' speaking development reported in community-member focus groups (Brunfaut & Green, 2017)

Benchmarking against international models. International assessments illustrate design principles that can inform local development. Since January 21, 2026, TOEFL iBT has used computer-delivered Listen and Repeat and

Take an Interview tasks for speaking, alongside an adaptive listening section comprising responses, conversations, announcements, and academic talks; score reports use a 1-6 scale aligned more directly with CEFR interpretation (ETS, n.d.-a, n.d.-b). IELTS uses a recorded face-to-face interview to elicit interaction across three parts, while PTE Academic demonstrates standardized technology-mediated delivery across multiple task types (British Council, n.d.; IELTS, n.d.; Pearson, n.d.-a, n.d.-b). Rather than importing these formats wholesale, Bangladesh’s tertiary sector can adapt their underlying principles: explicit specifications, sampling across task types, transparent rating criteria, rater training and monitoring, and clear reporting of score meaning.

Table 2

Illustrative comparison of oral-skill assessment components across selected tests

Test/model	Speaking elicitation	Listening elicitation	Design features useful for localization
TOEFL iBT	Computer-delivered Listen and Repeat + Take an Interview tasks	Adaptive mix of response, conversation, announcement, and academic-talk items	Concise standardized tasks; adaptive delivery; updated 1-6 score scale
IELTS	Face-to-face recorded interview (3 parts)	Recorded monologic + dialogic input	Interactional speaking; clear part structure; human examiner
PTE Academic	Computer-delivered multi-task battery	Multiple listening item types	Technology-mediated delivery; scalable administration

Test/model	Speaking elicitation	Listening elicitation	Design features useful for localization
Proposed tertiary framework (Bangladesh)	Hybrid: paired discussion + brief individual talk; optional integrated task	Short academic + campus-life inputs with locally familiar accents	CEFR-linked descriptors; rater training; low-cost recording; periodic audits

Note. Test descriptions are based on public test-provider documentation (British Council, n.d.; ETS, n.d.-a, n.d.-b; IELTS, n.d.; Pearson, n.d.-a, n.d.-b).

7. Discussion

The synthesis suggests that improving tertiary oral assessment in Bangladesh requires moving beyond isolated innovations toward coordinated systems that produce comparable interpretations. International tests show that standardization is achieved through specification-driven design, careful sampling, and sustained quality assurance rather than through a single “best” task type. At the same time, international models operate under resource conditions and governance structures that may not transfer directly. The proposed framework therefore emphasizes minimal viable standardization: shared constructs and rubrics, locally relevant tasks, and scalable procedures for training and moderation.

A key tension concerns authenticity versus feasibility. Highly authentic interactional tasks can be time-intensive and difficult to score consistently. However, authenticity is not an all-or-nothing property; tasks can be designed to approximate target language use within feasible time constraints, and validity arguments can make explicit what aspects of the target domain are represented and what is not (Bachman & Palmer, 2010; Weir, 2005). In Bangladesh, paired speaking tasks and short integrated tasks offer a pragmatic compromise that elicits interaction while reducing examiner burden.

The analysis also highlights the importance of fairness and the social dimension of language testing. Decisions about what counts as “good speaking” are not value-neutral; they embed assumptions about desirable language use and legitimate forms of English. A Global Englishes orientation supports a shift toward intelligibility, pragmatic appropriateness, and communicative effectiveness, which may be more equitable for Bangladeshi learners than strict adherence to native-speaker norms (Jenkins, 2015; McNamara & Roever, 2006). Fairness also requires attention to differential access to preparation resources, technology, and supportive learning environments.

Finally, the potential washback benefits of oral assessment depend on transparency and credibility. When students and teachers understand what is being assessed and how performance is judged, assessment can support targeted instruction and meaningful feedback. Conversely, opaque oral assessments may be experienced as subjective and may generate resistance. Therefore, the framework treats rubrics, benchmark performances, and moderation procedures as central to legitimacy.

8. Proposed Framework for Development and Standardization

The proposed framework adapts international good practice to Bangladesh’s tertiary context while prioritizing feasibility, fairness, and instructional usefulness. It is staged because the standardization of oral-skills assessment requires iterative development and evidence collection. Each stage produces documentation and artifacts—specifications, tasks, rubrics, training materials, and validation evidence—that can be shared across institutions to build sector-level coherence.

8.1 Stage 1: Define purpose, stakeholders, and constructs

Test development should begin by clarifying the primary uses of scores (e.g., course certification, placement, graduation, or exit standards) and identifying stakeholders affected by decisions. Construct definitions should be aligned with target-language-use domains such as academic seminars, group projects, presentations, office-hour consultations, and entry-level workplace interactions. Importantly, constructs should recognize the realities of English use in Bangladesh and the wider region, where intelligibility and pragmatic

effectiveness may be more relevant than narrow native-speaker norms (Jenkins, 2015). A practical approach is to develop local proficiency descriptors referenced to CEFR levels and illustrated by benchmark recordings (Council of Europe, 2001).

8.2 Stage 2: Create test specifications and blueprints

Specifications operationalize constructs into task requirements: input types, length, interaction patterns, permitted preparation time, timing, and scoring weights. A blueprint also clarifies how tasks sample subskills (e.g., global vs. detailed listening, inferencing, interactional management, and discourse organization) to support construct representation. Specifications should explicitly identify accommodation procedures, test security requirements, and minimal technical standards for audio equipment (Weir, 2005).

8.3 Stage 3: Design tasks and scoring instruments

Task design should balance authenticity with control. A feasible speaking model could combine (a) a paired discussion on a familiar academic or campus topic, (b) a short individual presentation or narrative, and (c) an optional integrated task in which students listen to a brief input and respond. This structure samples interactional competence while limiting examiner time. Listening assessment can combine short academic and semi-academic recordings with items that require meaning construction rather than surface word matching. Analytic speaking scales should distinguish key dimensions, such as pronunciation and intelligibility, grammatical and lexical resources, fluency, and interactional effectiveness. Such scales can support actionable feedback, but their descriptors require piloting for clarity and usability (Fulcher & Davidson, 2007; Luoma, 2004).

8.4 Stage 4: Pilot, analyze, and revise

Piloting should occur across multiple institutions to capture variability in student profiles and administration conditions. Listening tasks can be analyzed for facility and discrimination, while speaking ratings require systematic rater analyses. Where technical capacity exists, many-facet Rasch measurement can be used to examine rater severity, task difficulty, and scale functioning; where capacity is limited, simpler monitoring approaches (double marking of a sample, benchmark-based recalibration, and moderation meetings) can still

provide useful evidence (Fan & Knoch, 2019). Pilot results should be documented and used to revise tasks, rubrics, and administration instructions.

8.5 Stage 5: Standard setting and reporting

If institutions wish to make proficiency claims (e.g., “B1”), they must establish cut scores through a defensible standard-setting method. Performance-based methods that use benchmark performances linked to descriptors are often appropriate for speaking. Standard setting should be treated as an evidence-generating process, with explicit documentation of panel composition, training, and decision rules (American Educational Research Association et al., 2014; Brown & Abeywickrama, 2019). Score reports should communicate what scores mean in terms of expected communicative performance, not merely numerical totals.

8.6 Stage 6: Implementation with ongoing quality assurance

Standardization is sustained through rater training, calibration, and monitoring; periodic task review; and a continuous validation plan. Low-cost recording of speaking performances can support moderation, appeals, and rater feedback, while computer-delivered listening can stabilize delivery conditions. An annual review cycle should examine reliability evidence, rater monitoring, stakeholder feedback, and observed washback, and should document changes to specifications and tasks in an audit trail (American Educational Research Association et al., 2014; Kane, 2013).

Table 3
Illustrative validity evidence plans for tertiary listening and speaking assessment

Validity claim	Evidence type	Data sources	Quality assurance actions
Tasks represent target academic and workplace communication	Content evidence	Blueprint mapping; expert panels; student feedback	Revise specifications; document representativeness decisions

Validity claim	Evidence type	Data sources	Quality assurance actions
Scores reflect intended constructs, not irrelevant factors	Response-process evidence	Think-alouds; rater verbal protocols; technical checks	Standardize instructions; ensure audio quality; rater training
Ratings are consistent across raters and tasks	Internal structure evidence	Double-marking; rater monitoring; MFRM where feasible	Calibration sessions; benchmark recordings; drift checks
Scores relate to external indicators of proficiency	Relations-to-other-variables evidence	Course performance; external test correlations; teacher judgments	Investigate anomalies; refine constructs and reporting
Assessment supports equitable decisions and positive learning	Consequential/fairness evidence	Differential impact checks; stakeholder interviews; washback review	Accommodations; transparency measures; iterative policy revision

Note. Evidence categories follow contemporary validity theory and the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014; Kane, 2013).

9. Implications and Recommendations

National and institutional stakeholders can take several steps to strengthen validity and reliability while remaining sensitive to feasibility constraints.

First, tertiary institutions should adopt shared construct definitions and proficiency descriptors for listening and speaking, ideally referenced to CEFR levels but localized through stakeholder consultation and empirical review. A national working group or consortium of universities could coordinate this work and develop a shared repository of test specifications, tasks, and benchmark performances.

Second, assessment literacy development is essential. Professional development should cover construct definition, task design principles, rubric interpretation, standard setting, and basic reliability monitoring. Embedding assessment literacy in teacher education and ongoing institutional training would improve sustainability (Islam et al., 2021).

Third, rater systems should be treated as a core component of assessment infrastructure. Institutions should establish rater pools, provide initial and refresher training, calibrate raters using benchmark recordings, and monitor periodically for drift. Where capacity permits, many-facet Rasch analyses can provide additional diagnostic evidence concerning fairness (Fan & Knoch, 2019).

Fourth, technology should be used strategically. Standardized audio delivery for listening and routine recording of speaking can improve comparability and support moderation, but implementation must account for differential access and technical reliability. Clear minimum technical standards and contingency plans are necessary to prevent technology-related construct-irrelevant variance.

Fifth, transparency should be strengthened through documentation and reporting. Publishing rubrics, sample tasks, and score-interpretation guides can support fairness and positive washback. Institutions should also maintain an audit trail of changes to specifications and tasks, consistent with an ongoing validity argument (American Educational Research Association et al., 2014; Kane, 2013).

10. Limitations and Future Research

This paper is based on documentary and scoping evidence rather than on newly collected primary data from universities. As a result, conclusions about specific institutional practices should be treated as indicative rather than exhaustive. The analysis also reflects the uneven availability of publicly accessible test documents and detailed descriptions of tertiary assessment practices.

Future research should strengthen methodological triangulation by adding primary data such as interviews with university teachers, students, assessment

administrators, and employers, as well as classroom observations of assessment practices. Empirical piloting of proposed tasks and rubrics across multiple universities would enable more precise analysis of reliability, rater effects, and the relationship between oral assessment outcomes and academic performance. A sector-wide program of validation research, aligned to an explicit interpretive argument, would further strengthen the credibility and usability of standardized oral assessment in Bangladesh.

11. Conclusion

Listening and speaking assessment in Bangladeshi higher education requires systematic development and coordinated standardization to support communicative outcomes and equitable decisions. The evidence synthesized here indicates that construct ambiguity, limited authenticity in task design, scoring variability, and uneven institutional capacity undermine comparability and defensibility. A six-stage framework grounded in construct clarity, specification-driven task development, rater training and monitoring, standard setting, and continuing validation can strengthen score meaning and educational impact. By treating standardization as an ongoing capacity-building process rather than a single test product, Bangladeshi higher education can move toward fairer and more instructionally valuable assessment of oral skills.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115-129. <https://doi.org/10.1093/applin/14.2.115>
- Al Amin, M., & Greenwood, J. (2018). The examination system in Bangladesh and its impact: On curriculum, students, teachers and society. *Language Testing in Asia*, 8, Article 4. <https://doi.org/10.1186/s40468-018-0060-9>

- Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19-32. <https://doi.org/10.1080/1364557032000119616>
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. <https://doi.org/10.1191/1478088706qp063oa>
- British Council. (n.d.). *IELTS test format explained*. Retrieved August 1, 2026, from <https://takeielts.britishcouncil.org/take-ielts/test-format>
- Brunfaut, T., & Green, R. (2017). *English listening and speaking assessment in Bangladesh higher secondary schools: A baseline study*. British Council. <https://eprints.lancs.ac.uk/id/eprint/86353/>
- Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson.
- Buck, G. (2001). *Assessing listening*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732959>
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1-47. <https://doi.org/10.1093/applin/1.1.1>
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.
- Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge University Press.

- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.
- ETS. (n.d.-a). *TOEFL iBT speaking section*. Retrieved August 1, 2026, from <https://www.ets.org/toefl/test-takers/ibt/about/content/speaking.html>
- ETS. (n.d.-b). *TOEFL iBT test content and structure*. Retrieved August 1, 2026, from <https://www.ets.org/toefl/test-takers/ibt/about/content.html>
- Fan, J., & Knoch, U. (2019). Fairness in language assessment: What can the Rasch model offer? *Papers in Language Testing and Assessment*, 8(2), 117-142.
- Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Routledge.
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. Routledge.
- Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, 15(3), 219-236. <https://doi.org/10.1080/15434303.2018.1453816>
- Green, A. (2014). *Exploring language assessment and testing: Language in action*. Routledge.
- Hamid, M. O., & Baldauf, R. B., Jr. (2008). Will CLT bail out the bogged down ELT in Bangladesh? *English Today*, 24(3), 16-24. <https://doi.org/10.1017/S0266078408000254>
- Hamid, M. O., & Erling, E. J. (2016). English-in-education policy and planning in Bangladesh: A critical examination. In R. Kirkpatrick (Ed.), *English language education policy in Asia* (pp. 25-48). Springer.
- Hasan, M. M. (2020). Using various assessment and feedback tools for teaching English at the undergraduate level in Bangladesh. *International Journal of Science and Business*, 4(1), 214-227. <https://doi.org/10.5281/zenodo.3703145>
- IELTS. (n.d.). *IELTS Academic format: Speaking*. Retrieved August 1, 2026, from <https://ielts.org/take-a-test/test-types/ielts-academic-test/ielts-academic-format-speaking>

- Islam, M. S., Hasan, M. K., Rahman, M. M., Karim, A., Rahman, T. A., Seraj, P. M. I., & Singh, M. K. M. (2021). English language assessment in Bangladesh today: Principles, practices, and problems. *Language Testing in Asia*, 11, Article 2. <https://doi.org/10.1186/s40468-020-00116-z>
- Jenkins, J. (2015). *Global Englishes* (3rd ed.). Routledge.
- Kabir, M. S. (2015). Testing English at tertiary level: An overview. *Crossings: A Journal of English Studies*, 6(1), 196-213.
- 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/jedem.12000>
- Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27-48). Cambridge University Press.
- Levac, D., Colquhoun, H., & O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, 5, Article 69. <https://doi.org/10.1186/1748-5908-5-69>
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage.
- Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Blackwell.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). American Council on Education/Macmillan.
- Pearson. (n.d.-a). *PTE Academic test format: Listening*. Retrieved August 1, 2026, from <https://www.pearsonpte.com/pte-academic/test-format/listening>
- Pearson. (n.d.-b). *PTE Academic test format: Speaking & Writing*. Retrieved August 1, 2026, from <https://www.pearsonpte.com/pte-academic/test-format/speaking-writing>
- Rahman, M. M., & Pandian, A. (2018). A critical investigation of English language teaching in Bangladesh. *English Today*, 34(3), 43-49. <https://doi.org/10.1017/S026607841700061X>

- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., ... Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467-473. <https://doi.org/10.7326/M18-0850>
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.
- World Bank. (2019). *Bangladesh tertiary education sector review: Skills and innovation for growth*. <https://documents.worldbank.org/en/publication/documents-reports/documentdetail/303961553747212653>