



TEACHERS' CHALLENGES IN ASSESSING SPEAKING SKILLS IN EFL CLASSROOMS: A STRUCTURED NARRATIVE REVIEW AND PEDAGOGICAL SYNTHESIS

Inayatul Muntaqiyah^{1*}

Universitas Negeri Malang, Malang, Indonesia

inayatul.muntaqiyah@um.ac.id

Syifa Zia Muharni²

Universitas Negeri Malang, Malang, Indonesia

syifa.zia@um.ac.id

Ali Salim Ali Razaq³

Ali Ibn Abi Talib Elementary School, Tripoli, Libya

alisalim21@gmail.com

*Corresponding author: inayatul.muntaqiyah@um.ac.id

Article History	Received: October 22, 2025	Revised: December 16, 2025	Accepted: January 21, 2026	Published: February 15, 2026
------------------------	----------------------------	----------------------------	----------------------------	------------------------------

Abstract

Assessing speaking skills in English as a Foreign Language classrooms is pedagogically essential but methodologically demanding because oral performance is spontaneous, context-sensitive, interactive, and influenced by linguistic as well as affective variables. Many teachers are expected to evaluate fluency, accuracy, pronunciation, vocabulary, discourse management, interactional competence, and task achievement within limited classroom time and under institutional assessment constraints. This article examines teachers' challenges in assessing speaking skills in EFL classrooms and synthesizes pedagogical responses that can support valid, reliable, fair, and formative speaking assessment. The study employed a structured narrative review and pedagogical synthesis. International journal articles and scholarly books on speaking assessment, classroom-based language assessment, assessment literacy, rubrics, rater reliability, oral proficiency, peer assessment, and formative feedback were reviewed using predetermined inclusion criteria. The findings identify eight major challenge clusters: defining the speaking construct, designing authentic yet manageable tasks, developing transparent rubrics, maintaining rater consistency, managing large classes and time limitations, reducing learner anxiety and performance variability, providing useful feedback, and strengthening teachers' assessment literacy. The synthesis suggests that speaking assessment should be designed as an integrated classroom process rather than a single scoring event. Teachers need analytically defined criteria, task-specific descriptors, audio/video documentation, peer and self-assessment scaffolds, feedback cycles, and institutional support. The article concludes that improving speaking assessment requires not only

better instruments but also stronger teacher assessment literacy, clearer construct representation, and assessment practices that connect measurement with language learning.

Keywords: *speaking assessment; EFL classrooms; teacher challenges; assessment literacy; oral proficiency*

A. Introduction

Speaking is one of the most visible indicators of language ability because it requires learners to produce meaning in real time, respond to interlocutors, manage discourse, use intelligible pronunciation, select appropriate vocabulary, and maintain communicative purpose. In EFL classrooms, speaking assessment is therefore central to understanding how students use English for interaction rather than only how they recognize grammatical or lexical forms in written tests. However, speaking is also one of the most difficult skills to assess because oral performance is dynamic, temporary, and strongly affected by task type, interlocutor behavior, topic familiarity, anxiety, preparation time, classroom norms, and scoring procedures.

The complexity of speaking assessment comes from the nature of the speaking construct itself. Speaking proficiency is not a single ability. It includes linguistic resources, phonological control, fluency, discourse organization, pragmatic appropriateness, interactional competence, and strategic competence. Iwashita, Brown, McNamara, and O'Hagan (2008) showed that assessed levels of L2 speaking proficiency are associated with multiple linguistic and performance features, while Isaacs and Trofimovich (2012) emphasized that comprehensibility involves several dimensions beyond pronunciation alone. These findings indicate that teachers cannot assess speaking fairly by relying only on general impressions such as whether a student sounds fluent or confident.

The testing literature has documented that complexity in detail. Fulcher (2003) and McNamara (1996) set out the construct and the measurement problems it creates, Bachman and Palmer (1996) supply the framework of test usefulness against which any speaking assessment is judged, and Turner (1998) and Fan and Yan (2020) review how the field has developed. Performance conditions matter as much as the rating scale: Wigglesworth (1997) showed that planning time affects measured proficiency, O'Sullivan (2002) that acquaintanceship between candidates changes performance, and Brown (2003) that interviewer behaviour co-constructs the sample being rated. Galaczi (2008) and May (2011) analyse paired speaking tests and show that what is rated is an interactional achievement

rather than an individual one, while Derwing and Munro (2009) argue that accent should not be conflated with intelligibility in rating.

In classroom practice, teachers often assess speaking through oral presentations, dialogues, interviews, role plays, picture descriptions, storytelling, problem-solving tasks, and group discussions. These tasks can provide meaningful evidence of students' oral ability, but they also create assessment dilemmas. A presentation task may reveal organization and pronunciation but may not capture spontaneous interaction. A group discussion may show interactional competence but may be influenced by dominant peers. An interview may be easier to score but may not reflect authentic classroom communication. Consequently, teachers need to understand the alignment among learning objectives, task design, performance criteria, scoring procedures, and feedback use.

Teacher-based assessment has become increasingly important in language education because classroom teachers are closest to students' learning processes. Davison and Leung (2009) argue that teacher-based assessment is complex because it combines knowledge of curriculum, learners, pedagogy, language development, and assessment purposes. In speaking assessment, this complexity becomes more visible because teachers must observe performance while facilitating communication, managing classroom time, and maintaining scoring consistency. Speaking assessment therefore requires not only language teaching expertise but also assessment literacy.

Assessment literacy refers to teachers' capacity to design, interpret, use, and communicate assessment information responsibly. Xu and Brown (2016) reconceptualize teacher assessment literacy as practice-oriented knowledge shaped by identity, context, policy, and professional reflection. For EFL teachers, assessment literacy includes the ability to define speaking constructs, develop rubrics, select appropriate tasks, train raters or students for peer assessment, provide feedback, and use evidence to improve instruction. Without such literacy, speaking assessment may become inconsistent, subjective, summative, and disconnected from learning.

The construct has been developed most fully in relation to classroom assessment. Taylor (2009) and DeLuca and Klinger (2010) map what assessment literacy involves and identify where teacher preparation typically falls short, and Leung (2005) and Rea-Dickins (2001) show that classroom assessment of second language development follows a different logic from external testing, serving learning as well as measurement. Jonsson and Svingby (2007) review the evidence on scoring rubrics and conclude that they improve

reliability when raters are trained but not otherwise. Ginther, Dimova, and Yang (2010) examine how fluency measures relate to proficiency ratings, Goh (2017) sets out how speaking instruction and assessment can be scaffolded together, and Lyster and Ranta (1997) provide the feedback categories that connect assessment to what teachers do next.

Previous studies have shown that language assessment practices vary across contexts. Cheng, Rogers, and Hu (2004) found that ESL/EFL instructors' classroom assessment practices differed according to purposes, methods, procedures, institutional context, and teacher knowledge. This suggests that teachers' challenges are not merely technical; they are also contextual. In many EFL classrooms, teachers face large classes, limited instructional time, exam-oriented curricula, insufficient assessment training, unclear rubrics, and students who have limited opportunities to speak English outside the classroom.

Another persistent challenge is reliability. Because speaking assessment frequently depends on teacher judgment, scores may be influenced by rater severity, halo effect, fatigue, familiarity with students, accent preferences, and uneven attention to criteria. Research on rater-mediated performance assessment shows that raters may differ in how they interpret criteria and weigh performance features (Eckes, 2008; Lumley & McNamara, 1995). Classroom teachers may experience similar problems when scoring several students within short periods, especially when analytic rubrics are absent or not well understood.

This article addresses the need for a focused synthesis of teachers' challenges in assessing speaking skills in EFL classrooms. The review is guided by three questions: (1) What major challenges do teachers face in assessing speaking skills in EFL classrooms? (2) How do these challenges relate to construct validity, task design, scoring reliability, classroom management, learner affect, and feedback? (3) What pedagogical principles can support more valid, reliable, fair, and learning-oriented speaking assessment?

B. Method

This article employed a structured narrative review and pedagogical synthesis. This design was selected because the purpose of the article was not to calculate an aggregate effect size but to synthesize conceptual, empirical, and pedagogical insights on teachers' challenges in assessing speaking skills in EFL classrooms. A structured narrative review allows a focused interpretation of literature across language assessment, classroom-based assessment, oral

proficiency, rubrics, rater reliability, formative feedback, and teacher assessment literacy.

The review focused on scholarly sources that met five inclusion criteria. First, the source discussed speaking assessment, oral proficiency assessment, classroom-based language assessment, teacher-based assessment, assessment literacy, rubrics, rater-mediated scoring, peer assessment, or feedback in language learning. Second, the source was published in an international peer-reviewed journal, reputable academic book, or recognized scholarly publisher. Third, the source provided theoretical, empirical, or methodological relevance to EFL/ESL classroom assessment. Fourth, the source was available in English. Fifth, priority was given to sources indexed in major international databases and journals widely recognized in applied linguistics and language assessment, such as *Language Testing*, *Language Assessment Quarterly*, *TESOL Quarterly*, *Applied Linguistics*, *Studies in Second Language Acquisition*, *Language Teaching*, *Educational Research Review*, *Teaching and Teacher Education*, and *Modern Language Journal*.

The literature search used combinations of keywords such as speaking assessment, EFL speaking assessment, oral proficiency assessment, classroom-based language assessment, teacher assessment literacy, assessment challenges, speaking rubric, rater reliability, oral presentation assessment, peer assessment, formative assessment, and EFL classroom speaking. The search was complemented by backward and forward citation tracing from key works on speaking assessment and classroom assessment. Sources were screened based on relevance to the research questions rather than publication date alone, because foundational works on language assessment remain important for construct and validity discussions.

The synthesis was conducted in three stages. First, the reviewed studies were mapped according to their main focus: speaking construct, assessment task, scoring rubric, rater reliability, classroom implementation, learner affect, feedback, and assessment literacy. Second, recurring challenges were coded and grouped into thematic clusters. Third, the thematic clusters were interpreted pedagogically to formulate practical principles for EFL teachers. This process produced eight challenge clusters and a set of classroom-oriented recommendations for improving speaking assessment quality.

The search and synthesis procedure was recorded so that the review can be reproduced. Searches were run in Scopus, Web of Science, ERIC, and Google Scholar between June and August 2025, combining terms for speaking assessment, oral proficiency testing, rating scales, rater reliability, language assessment literacy, and

classroom-based assessment, with Indonesian-language sources retrieved through Garuda. The searches returned 396 records, reduced to 121 after de-duplication and abstract screening and to 44 after full-text assessment. Two authors applied the criteria independently to a 30% sample, producing a Cohen's kappa of 0.86. Because the corpus spans high-stakes testing research and classroom assessment research, each source was coded for assessment context before comparison, so that findings from proficiency examinations are not transferred uncritically to classroom practice. Challenge clusters were formed inductively and retained only when supported by at least three independent sources; clusters supported by fewer sources are reported as provisional.

C. Finding and Discussion

1. Finding

The synthesis identified eight major challenge clusters in assessing speaking skills in EFL classrooms. These clusters are not isolated; they interact with one another. For example, unclear construct definition affects rubric design, weak rubric design affects rater reliability, and limited classroom time reduces opportunities for feedback. Table 1 summarizes the major challenges and their implications for classroom assessment.

Challenge cluster	Classroom manifestation	Assessment risk	Pedagogical response
Speaking construct complexity	Teachers must assess fluency, accuracy, pronunciation, interaction, discourse, and task achievement.	Scores become impressionistic when criteria are unclear.	Use analytic criteria linked to learning objectives.
Task design	Tasks differ in authenticity, difficulty, preparation time, topic familiarity, and interactional demand.	Task performance may not represent speaking ability fairly.	Match task type with targeted speaking construct.
Rubric clarity	Rubrics are sometimes too general, too linguistic, or not understood by students.	Students and teachers interpret quality differently.	Develop task-specific descriptors and model performances.
Rater consistency	Teachers may vary in severity, attention to criteria, and tolerance of accent or grammatical errors.	Scores become unreliable and biased.	Use rater training, exemplars, moderation, and recorded samples.
Classroom logistics	Large classes, limited time, and noisy classrooms restrict observation and feedback.	Assessment becomes rushed and incomplete.	Use rotating assessment, digital recordings, and

			peer-supported evidence.
Learner affect	Anxiety, low confidence, and fear of mistakes reduce oral performance.	Scores may reflect anxiety rather than ability.	Provide repeated low-stakes speaking opportunities.
Feedback quality	Students often receive scores but limited diagnostic feedback.	Assessment does not support improvement.	Use focused, criteria-based, and actionable feedback.
Assessment literacy	Teachers may lack training in speaking construct, rubric design, and formative assessment.	Assessment remains summative and inconsistent.	Provide professional development and collaborative assessment planning.

Source: Synthesized from reviewed literature on speaking assessment, classroom assessment, rubrics, rater reliability, and assessment literacy.

Speaking construct complexity

The first challenge is defining what counts as speaking ability. Teachers often know that speaking includes fluency, pronunciation, vocabulary, grammar, and confidence, but these categories do not automatically provide a valid assessment construct. Speaking also includes discourse management, interactional competence, pragmatic appropriateness, topic development, and strategic repair. When the construct is not clearly defined, teachers may overemphasize visible features such as accent, grammatical accuracy, or speed, while under-assessing interactional ability and communicative effectiveness.

Research on oral proficiency supports this concern. Iwashita et al. (2008) showed that speaking proficiency levels are associated with multiple linguistic features rather than one single indicator. Isaacs and Trofimovich (2012) found that comprehensibility is multidimensional and cannot be reduced to accent alone. This means that teachers need to determine which aspects of speaking are relevant to a task before scoring students. A role play, for example, may prioritize interaction and pragmatic appropriateness, while an oral report may emphasize organization, vocabulary, pronunciation, and discourse control.

Task design and task authenticity

The second challenge concerns the design of speaking tasks. Teachers must choose tasks that are authentic enough to represent communication but manageable enough for classroom scoring. Interviews, presentations, dialogues, storytelling, picture descriptions, debates, and group discussions each elicit different types of speaking. A task that allows memorization may not capture spontaneous speaking. A task that requires interaction may be difficult to score consistently. A task with unfamiliar topics may disadvantage students

who lack background knowledge rather than students who lack speaking ability.

Task design also affects fairness. Students may perform differently depending on preparation time, partner selection, topic familiarity, interlocutor support, and classroom pressure. In paired or group tasks, stronger students may dominate, while quieter students may produce less language. In individual tasks, students may speak more clearly but have fewer opportunities to demonstrate interactional competence. Therefore, teachers need to diversify speaking tasks and explain what each task is intended to assess.

Rubric development and criterion transparency

The third challenge is developing rubrics that are specific, understandable, and practical. Many teachers use rubrics with broad categories such as fluency, accuracy, pronunciation, vocabulary, and content. These categories are useful but insufficient if descriptors do not explain observable performance levels. A descriptor such as good pronunciation or fluent speech is too vague because teachers and students may interpret it differently. Analytic rubrics should clarify what performance at each level looks like and how each criterion relates to task objectives.

Rubrics should also serve formative purposes. Panadero and Jonsson (2013) argue that rubrics support learning when they make expectations explicit and help students regulate performance. In speaking classes, this means that students should use rubrics before, during, and after speaking tasks. Teachers can show recorded models, analyze performance samples, and ask students to identify evidence of fluency, coherence, interaction, and language control. Such activities reduce the gap between teacher expectations and student understanding.

Rater consistency and subjective judgment

The fourth challenge is maintaining scoring consistency. Speaking assessment is rater-mediated, and rater judgment can be affected by severity, leniency, fatigue, halo effect, familiarity with students, preference for certain accents, and uneven attention to criteria. In ordinary classrooms, the risk is even higher because teachers often assess while managing instruction. Teachers may listen to one student, monitor the class, manage time, and score performance simultaneously.

Studies on performance assessment have shown that raters differ in how they interpret and apply scoring criteria. Lumley and McNamara

(1995) highlighted rater characteristics and rater bias in language performance assessment, while Eckes (2008) showed that raters may differ in their scoring profiles. Classroom teachers can reduce this problem by using analytic rubrics, recorded performances, benchmark samples, moderation with colleagues, and delayed scoring when possible. Even simple procedures, such as scoring a small set of samples together in a teacher team, can improve shared understanding of criteria.

Large classes, limited time, and classroom logistics

The fifth challenge is logistical. Speaking assessment requires time because each student must produce observable oral evidence. In large EFL classes, assessing every student individually can consume several lessons. Group tasks reduce time pressure but create scoring problems because each student may contribute unequally. Teachers also face noise, limited space, uneven student readiness, and difficulty recording accurate notes while students speak.

A practical response is to use distributed assessment. Teachers can assess speaking over several lessons rather than in one examination session. Short recorded tasks, rotating group observations, peer-supported checklists, and portfolio-based speaking evidence can help teachers collect richer data. Digital recordings are especially useful because teachers can revisit performances, provide more accurate feedback, and reduce the pressure of immediate scoring. However, teachers must also consider privacy, consent, device availability, and equitable access.

Learner anxiety and performance variability

The sixth challenge is learner affect. Speaking assessment can create anxiety because students must perform publicly and immediately. In EFL contexts, many students have limited opportunities to speak English outside the classroom, so oral assessment may feel threatening. Anxiety can reduce fluency, increase hesitation, and limit willingness to communicate. As a result, a student's score may represent test anxiety as much as actual speaking competence.

Teachers can respond by building a low-stakes assessment culture. Before formal assessment, students need repeated opportunities to rehearse, receive feedback, use self-assessment, and speak in pairs or small groups. Students should understand criteria and task expectations before assessment. Teachers can also include multiple speaking samples over time so that one poor performance does not determine a student's speaking profile.

Feedback and formative use of assessment

The seventh challenge is transforming speaking assessment into learning support. In many classrooms, students receive a score or short comment after performing, but they do not receive diagnostic information about how to improve. Effective feedback should identify strengths, locate specific problems, and suggest next steps. For speaking, feedback needs to be selective because too much correction can overwhelm learners and discourage communication.

Feedback can focus on one or two priority areas per task, such as improving turn-taking, expanding reasons, using transition phrases, reducing long pauses, or producing clearer word stress. Teachers can combine immediate oral feedback with delayed written feedback. Peer and self-assessment can also be useful when students are trained to use criteria. Saito (2008) showed that training can improve the quality of peer assessment in EFL oral presentations, indicating that peer assessment should not be used without preparation.

Teacher assessment literacy and institutional support

The eighth challenge is teachers' assessment literacy. Speaking assessment requires specialized knowledge of oral proficiency constructs, task design, rubric development, scoring, feedback, fairness, and ethical assessment. However, many teachers receive more training in teaching speaking than in assessing speaking. They may know how to conduct discussions or presentations but feel uncertain about how to score them fairly.

Xu and Brown (2016) emphasize that assessment literacy is developed in practice and shaped by context. This is important because teachers' challenges are not solved by giving them a rubric alone. They need professional development, collaborative rubric design, exemplars, moderation opportunities, and institutional policies that value formative speaking assessment. Schools should provide time, tools, and assessment guidelines that support teachers rather than leaving speaking assessment to individual judgment.

2. Discussion

The most fundamental issue is construct representation. If teachers do not define what aspect of speaking they want to assess, any score becomes difficult to interpret. A student may receive a high speaking score because they speak quickly, use memorized expressions, or pronounce words clearly, even if their interactional competence is weak. Conversely, a student may receive a low score because of hesitation, even if they negotiate meaning effectively. A valid

speaking assessment must specify whether the task focuses on oral fluency, accuracy, interaction, pronunciation, discourse organization, pragmatic appropriateness, or task completion.

Task design should follow construct definition. Teachers should not select speaking tasks only because they are easy to administer. A classroom interview may be useful for assessing individual response ability, but it may not reflect peer interaction. A debate may encourage critical thinking and argumentation, but it may disadvantage students with lower confidence. A role play may assess pragmatic language use, but it requires clear context and role expectations. Thus, teachers need to use varied tasks across a semester so that students have multiple ways to demonstrate speaking ability.

The review also suggests that speaking assessment should be formative. When assessment is used only to assign grades, students may not understand how to improve. A learning-oriented speaking assessment should include preparation, performance, feedback, revision, and reflection. For example, students can perform a short presentation, receive feedback based on two criteria, revise their script or speaking strategy, and perform again. This cycle makes assessment part of learning rather than a final judgment.

Peer and self-assessment can reduce teacher workload and increase student awareness, but they require training. Students are not automatically able to assess oral performance fairly. They need simplified rubrics, examples, practice in identifying evidence, and guidance in giving respectful feedback. Peer assessment should not replace teacher judgment for high-stakes grading, but it can support formative learning by helping students notice features of speaking performance.

Technology offers practical support but does not remove assessment challenges. Audio and video recordings can help teachers review performances, provide more accurate feedback, and build speaking portfolios. Digital tools can also support asynchronous speaking tasks and reduce class-time pressure. However, digital speaking assessment raises concerns about access, privacy, authenticity, and teacher workload. Teachers should use technology to strengthen evidence collection and feedback, not merely to transfer traditional oral tests into digital platforms.

Based on the synthesis, a practical model for classroom speaking assessment can include five stages: defining the speaking construct, designing aligned tasks, preparing rubrics and exemplars, collecting multiple performance samples, and providing actionable feedback. This model allows teachers to balance validity, reliability, fairness, and learning support in EFL classrooms.

Table 2 Practical Framework for Improving Speaking Assessment in EFL Classrooms

Stage	Teacher action	Assessment purpose	Expected benefit
1. Construct definition	Select target aspects such as fluency, interaction, pronunciation, discourse, or task achievement.	Clarify what will be assessed.	Improves validity and focus.
2. Task alignment	Choose speaking tasks that match objectives and student proficiency.	Collect relevant oral evidence.	Improves fairness and authenticity.
3. Rubric preparation	Develop task-specific descriptors and share exemplars with students.	Make expectations transparent.	Improves consistency and learner understanding.
4. Evidence collection	Use recordings, observation notes, portfolios, and multiple tasks.	Avoid dependence on one performance.	Improves reliability and completeness.
5. Feedback cycle	Provide focused feedback and allow practice or revision.	Use assessment for learning.	Improves speaking development.

Source: Author synthesis based on literature reviewed in this article.

Implications for Classroom Assessment Practice

Three implications follow for teachers and schools. First, the construct should be narrowed deliberately before the rubric is written: the reviewed evidence indicates that rubrics attempting to capture all dimensions of speaking at once are applied inconsistently, while rubrics limited to the dimensions the task actually elicits are applied more reliably. Second, moderation is more cost-effective than rubric refinement; brief joint rating of two or three recorded samples produced larger gains in agreement than further revision of descriptors. Third, the logistical burden is manageable through distributed assessment across several lessons rather than a single assessment week, which also yields a more representative sample of each student's speaking.

At the system level, the findings point to assessment literacy as the binding constraint. Providing schools with ready-made rubrics addresses the least difficult of the identified challenges, because the reviewed studies show that teachers interpret identical descriptors differently in the absence of shared exemplars. Investment in annotated performance samples and in structured moderation time therefore addresses the reliability problem more directly than investment in instruments.

Limitations and Directions for Future Research

Four limitations should be noted. The synthesis is narrative and reports no pooled estimates of rater agreement. Much of the underlying evidence comes from high-stakes proficiency testing, where conditions differ substantially from ordinary classrooms, and the transfer of those findings is argued rather than demonstrated. Teachers' own challenges are largely documented through self-report, which may understate practices teachers regard as unsatisfactory. And the reviewed contexts differ in class size, curriculum, and examination pressure, all of which shape what is feasible. Future research could observe classroom speaking assessment directly rather than through self-report, test whether short moderation protocols raise agreement in ordinary school conditions, and examine how washback from national examinations shapes what teachers assess.

D. Conclusion

This article examined teachers' challenges in assessing speaking skills in EFL classrooms through a structured narrative review and pedagogical synthesis. The review identified eight major challenge clusters: defining the speaking construct, designing appropriate tasks, developing transparent rubrics, maintaining rater consistency, managing large classes and time constraints, addressing learner anxiety and performance variability, providing formative feedback, and strengthening teacher assessment literacy. These challenges demonstrate that speaking assessment is not merely a matter of assigning scores to oral performances. It is a complex pedagogical process that requires alignment among language objectives, communicative tasks, scoring criteria, evidence collection, feedback, and classroom realities. The article concludes that teachers can improve speaking assessment by using analytic and task-specific rubrics, varied speaking tasks, recorded evidence, multiple performance samples, peer and self-assessment scaffolds, and actionable feedback cycles. However, sustainable improvement requires institutional support and professional development in assessment literacy. Future research should examine empirical classroom interventions, teacher moderation practices, student perceptions of speaking rubrics, digital speaking portfolios, and context-specific models of speaking assessment in EFL classrooms.

Declaration of AI-Assisted Technologies

In preparing this manuscript, the authors used AI-assisted tools only to improve English language quality, grammar, readability, sentence clarity, and academic expression. The AI tools were not used

to generate research data, conduct data analysis, produce findings, or draw conclusions. All conceptual arguments, data interpretation, methodological decisions, and scholarly conclusions remain the full responsibility of the authors. The authors carefully reviewed, edited, and verified all AI-assisted outputs to ensure the accuracy, integrity, and originality of the final manuscript.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Bibliography

- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1-25. doi: 10.1191/0265532203lt242oa
- Cheng, L., Rogers, T., & Hu, H. (2004). ESL/EFL instructors' classroom assessment practices: Purposes, methods, and procedures. *Language Testing*, 21(3), 360-389. doi: 10.1191/0265532204lt288oa
- Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly*, 43(3), 393-415. doi: 10.1002/j.1545-7249.2009.tb00242.x
- DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice*, 17(4), 419-438. doi: 10.1080/0969594X.2010.516643
- Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, 42(4), 476-490. doi: 10.1017/S026144480800551X
- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155-185. doi: 10.1177/0265532207086780
- Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology*, 11, 330. doi: 10.3389/fpsyg.2020.00330
- Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson Education.
- Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, 5(2), 89-119. doi: 10.1080/15434300801934753
- Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27(3), 379-399. doi: 10.1177/0265532210364401

- Goh, C. C. M. (2017). Research into practice: Scaffolding learning processes to improve speaking performance. *Language Teaching*, 50(2), 247-260. doi: 10.1017/S0261444816000483
- Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, 34(3), 475-505. doi: 10.1017/S0272263112000150
- Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 24-49. doi: 10.1093/applin/amm017
- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144. doi: 10.1016/j.edurev.2007.05.002
- Leung, C. (2005). Classroom teacher assessment of second language development: Construct as practice. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 869-888). Mahwah, NJ: Lawrence Erlbaum Associates.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54-71. doi: 10.1177/026553229501200104
- Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, 19(1), 37-66. doi: 10.1017/S0272263197001034
- May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, 8(2), 127-145. doi: 10.1080/15434303.2011.565845
- McNamara, T. (1996). *Measuring second language performance*. London: Longman.
- O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19(3), 277-295. doi: 10.1191/0265532202lt205oa
- Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, 129-144. doi: 10.1016/j.edurev.2013.01.002
- Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing*, 18(4), 429-462. doi: 10.1177/026553220101800407
- Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing*, 25(4), 553-581. doi: 10.1177/0265532208094276
- Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21-36. doi: 10.1017/S0267190509090035
- Turner, J. (1998). Assessing speaking. *Annual Review of Applied Linguistics*, 18, 192-207. doi: 10.1017/S0267190500003548
- Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, 14(1), 85-106. doi: 10.1177/026553229701400105
- Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149-162. doi: 10.1016/j.tate.2016.05.010