Criteria and Challenges for Rating Speaking Test Performance at the University of El Oued, Algeria

Aissa Berregui[1,*], Mohammed Naoua[2]

[1] Laboratory of Pragmatics and Discourse Analysis, English Language Department, Faculty of Arts and Languages, University of El Oued (Algeria)

[2] Laboratory of Pragmatics and Discourse Analysis, English Language Department, Faculty of Arts and Languages, University of El Oued (Algeria)

Date of submission: 12/06/2026 *Date of acceptance: 13/09/2026* *Date of publication: 10/10/2026*

Abstract:

Testing speaking is a problematic issue in language assessment, especially in English as foreign language (EFL) settings where speaking performance is scored by human raters. This study analyses the scoring of speaking tests at the University of El Oued EFL Department. The focus is on teachers' scoring criteria and the challenges influencing the rating process. The researchers adopted a qualitative descriptive case-study design. Data were gathered using non-participant observation of speaking test sessions and semi-structured interviews with eight EFL teachers involved in rating speaking ability. Observation recorded task format, topic choice, test time, teacher roles, scoring practices, recording, and moderated rating. Interviews investigated teachers' perceptions of assessment criteria, consistency, subjectivity, fairness and suggested improvements. The data were subjected to thematic analysis. The findings reveal inconsistent implementation of criteria and accentuate the need for promoting standardized rubrics, rater training, moderation, and stronger institutional support.

***Key words**: rater variability, measurement challenges, scoring criteria, speaking tests*

1. Introduction

Speaking is one of the most important language skills. It holds a central position in teaching and learning of English as Foreign Language (EFL) because the ability to speak a language reflects learners' level of oral communicative competence (Fulcher, 2003). In higher education contexts, students are required not only to develop linguistic knowledge but also to communicate orally in English for classroom discourse, oral presentation, and future workplace communication. This conceptualization of speaking is supported by communicative approaches to language testing and assessment, which view language ability as the capacity to use language appropriately and meaningfully in context rather than simply demonstrate a strong command of isolated grammatical forms and structures (Bachman, 1990; Bachman & Palmer, 1996; Canale & Swain, 1980). On this basis, Taylor (2011) asserts that the assessment of speaking skill has been recognized as an indispensable component of EFL education. Nevertheless, measuring speaking ability remains one of the most challenging and complex areas of language testing practices because speaking performance is real-time, interactive, context-sensitive and dynamic (Luoma, 2004).

In contrast to receptive language skills, during speaking, learners are required to mobilize various abilities simultaneously. Speakers need to retrieve vocabulary items, use grammar accurately and appropriately, control pronunciation, structure ideas logically, and manage real-time communicative demands (Bygate, 1987). The multidimensional nature of oral language performance and its complexity makes it difficult to measure objectively. Speaking ability encompasses not only linguistic correctness but also discourse management, interactional ability,

* Aissa Berregui.

strategic competence and the effective communication of meaning (Celce-Murcia et al., 1995). According to Luoma (2004), assessors of speaking should pay careful attention to the construct of speaking being measured, speaking tasks, rating scales, scoring rubrics, and the interpretation of oral discourse. Likewise, Fulcher (2003) affirms that speaking assessment is highly demanding because spoken productions are inherently transient, variable, and context-dependent.

As far as language assessment is concerned, the testing of speaking ability is generally considered as one form of performance assessment, where test-takers are evaluated according to their observable spoken performance. This is significant because it serves as a direct evidence of examinees' oral English ability, but it also complicates the measurement process since scoring is largely dependent on human judgment. In this regard, examiners are required to listen to and interpret oral performances, apply assessment criteria, and translate complex oral performance into scores. Additionally, speaking measurement is shaped by the following: the interaction with the test taker, the test task, the rater. More precisely, scores of speaking tests are not assigned based on learner ability alone; instead they are also affected by task characteristics, rating procedures, test conditions, rater interpretation (McNamara, 1996). In a similar vein, In'nami and Koizumi (2016) state that research in the field of second language speaking measurement has proved that factors such as task and rater aspects can cause score variability.

Scoring speaking tests is a very sensitive process due to the fact that raters commonly attend to various dimensions (criteria) of oral ability at the same time. These criteria can involve pronunciation, fluency, accuracy, lexical range, discourse coherence, cohesion, topical relevance, interactional competence, and task completion (Brown, 2004; Hughes, 2003). The main challenge, however, does not concern merely the identification of the aforementioned criteria. A more critical issue is related to the way raters interpret, prioritize, and employ them during scoring sessions (Upshur & Turner, 1995). In the absence of clearly defined assessment criteria, raters may base their decisions on general impression, background knowledge, or intuitive judgment (Wigglesworth, 1993). These practices may result in scoring inconsistency and give rise to questions about test validity and reliability (Eckes, 2005, Kim, 2015).

The most central concerns in speaking ability measurement are validity and reliability. The former is concerned with the extent to which the interpretation of scores is appropriate and meaningful, while the latter refers to the dependability and consistency of scores (Messick, 1989; Weir, 2005). When measuring speaking, reliability can be compromised by several factors including rater leniency, severity, tiredness, vague descriptors, short oral performance, topic differences, and time restrictions (Luoma, 2004; Wigglesworth, 1993). Validity can also be threatened when rating criteria do not fully reflect the construct of spoken language ability or when test scores are affected by construct-irrelevant variances including test stress, topic difficulty and familiarity, or the examiner's previous knowledge of the candidate (Khabbazbashi, 2017). Additionally, a test of speaking ability can be well-constructed in terms of task design, but it may yield problematic scores in case of unclear or inconsistent rating procedure (Bachman, 1990; McNamara, 1996).

1.1. Statement of the Problem

Within the Algerian EFL higher education context, the previously mentioned issues are highly required and important because educators frequently perform their assessment duties under certain institutional, pedagogical, and practical constraints. With reference to previous studies conducted in Algerian universities, researchers have reported some problems concerning speaking assessment criteria, the use of scoring rubrics, time allotment, rater-based scoring methods (Sadoune, 2020; Senouci, 2024). Furthermore, despite the fact that testing speaking has been largely tackled in research on language assessment, there is a dearth of studies on the way students' performance in speaking tests is rated in Algerian higher education EFL context. This

research gap is particularly relevant to the university of El Oued EFL department, where the researchers' observation of testing speaking skills sessions revealed a range of practices and issues related to test administration, rating criteria, scoring consistency, and rater roles. These practices may influence rating quality, and thereby compromising oral test validity and reliability. This reality led the researchers to think that student' speaking test scores may be based not only on their oral language proficiency but also on procedural and scorer- related aspects.

1.2. Aims of the Study

This research is an endeavor to investigate the rating of speaking test performance in EFL department at the University of El Oued in Algeria. The specific aims of the research are:

- *To identify the rating criteria used by EFL teachers at the University of El Oued when scoring learners' speaking test performance.*
- *To examine how scoring is conducted during speaking test sessions.*
- *To explore the difficulties encountered by EFL teachers when rating speaking tests.*
- *To suggest possible procedures for reinforcing reliability of speaking tests in this context.*

1.3. Research Questions

The study addresses the following research questions:

- *Which rating criteria are used at the University of El Oued when scoring speaking test performance?*
- *How is scoring conducted during speaking test sessions?*
- *What difficulties are encountered when rating learners' speaking test performance?*
- *How could the reliability of speaking tests be reinforced in this context?*

1.4. Research Hypotheses

In light of the research questions, the following hypotheses are formulated:

- ✓ *Reliance on individually selected rating criteria may produce inconsistent judgment of learners' speaking test performance.*
- ✓ *Scoring speaking tests without standardized procedures would increase rater subjectivity.*
- ✓ *Procedural constraints and rater-related challenges are likely to affect the fairness of speaking test scores.*
- ✓ *The use of rating scales, rater training and standardized assessment procedures could reinforce speaking test reliability.*

1.5. Significance of the Study

This study is significant as it could contribute to enriching research on language assessment and enhancing pedagogical practices in the Algerian EFL context. Theoretically, it broadens existing research on testing speaking, rater variability, validity, reliability, and rating criteria in EFL University settings. Practically, the study can assist educators, curriculum designers, and academic administrators in identifying the gaps and weaknesses in current practices for rating speaking tests and establishing more reliable and systematic procedure for rating this kind of performance tests.

2. Literature review

2.1. Overview of Speaking Skill Assessment in EFL Settings

The testing of speaking ability is a central area within the field of language assessment because it offers a concrete evidence of students' ability to use oral language for communication. The rise of communicative language testing shifted the emphasis from the mastery of language forms to learners' ability to use language appropriately in communicative contexts (Fulcher & Davidson, 2007). Therefore, the ultimate aim of speaking tests is to assess students' ability in expressing ideas, organizing spoken discourse, communicating with others, and reacting appropriately to communicative demands (Bygate, 1987; Luoma, 2004).

There is a close relationship between communicative competence and the construct of spoken language ability. Canale and Swain (1980) view that the components of communicative

competence are: grammatical, sociolinguistic, discourse, and strategic competencies. This framework is very beneficial when it comes to measuring speaking because effective spoken communication requires more than grammatical accuracy. Learners are expected to structure their ideas, use appropriate language, reply to interlocutors, and handle communication problems strategically (Canale, 1983). Similarly, Bachman (1990) introduced a model of communicative language ability, which reinforces the claim that speaking ability includes both knowledge of language and the ability to use it strategically in communicative contexts.

Speaking assessment can be done for formative or summative purposes. The aim of formative assessment is supporting the learning process through feedback provision, and determining areas of improvement, whereas the purpose of summative assessment is evaluating learners' achievements at the end of a semester or instructional cycle (Green, 2014). In university EFL contexts, speaking tests used for summative assessment are very important and decisive because they contribute to students' final grades and academic advancement. Accordingly, during summative speaking assessment, examiners should pay careful attention when selecting the scoring criteria and designing speaking test tasks (Hughes, 2003).

The design and type of speaking tasks have a major impact on the quality and strength of evidence obtained from assessment. Oral interviews, discussions, presentations, role plays, picture descriptions, and paired interactions do not elicit the same aspects of spoken language ability. To illustrate, paired tasks may demonstrate interactional competence, meaning negotiation, and turn-taking strategies, while presentations can emphasize organization and extended spoken discourse (Galaczi, 2014; Lazaraton, 1996). As a result, the choice of test tasks can affect both the students' performance during the test and the rating of this performance (Iwashita et al., 2008).

In the Algerian EFL context, the testing of speaking is more complicated because students have limited opportunities to speak English outside the classroom. This may cause students' test anxiety, which has been proved to influence cognitive processing and spoken production, particularly in case of formal speaking tests (Horwitz et al., 1986; MacIntyre & Gardner, 1994). These considerations may obscure the distinction between students' real speaking ability and their temporary test performance. Therefore, distinguishing between competence and performance is crucial in testing speaking because speech is generated under instantaneous psychological and situational pressure (Skehan, 1998).

2.2. Rating Criteria for Measuring Speaking Performance

Rating criteria constitute the basis for measuring speaking performance, as they outline the features of speaking ability that should be observed and assessed by raters. The most widely used criteria comprise pronunciation, fluency, vocabulary, accuracy, coherence, cohesion, content relevance, interactional ability, and task completion (Brown, 2004; Luoma, 2004). These criteria capture the multifaceted nature of spoken language ability and thus enable raters to transform a multi-component speaking performance into meaningful scores (Fulcher, 2003).

One of the most frequently cited indicators of speaking ability is the criteria of fluency. It is commonly associated with speech continuity, smoothness, and appropriate pace, but it must not be mixed with speed. A large body of research on foreign language oral performance tests considers fluency, accuracy, and complexity as the main aspects of speaking proficiency (Segalowitz, 2010; Tavakoli & Skehan, 2005). While testing speaking, fluency should be examined with reference to pausing patterns, hesitation, self-repair, speech continuity rather than rapidity alone (Luoma, 2004).

Pronunciation is also of equal importance in testing speaking because it influences both comprehensibility and intelligibility. As per Derwing and Munro (2005), a distinction should be made among intelligibility, accent, and comprehensibility because non-native accented speech should not be considered as unintelligible (difficult to understand). Such a distinction is of utmost significance especially in EFL contexts, where examiners should not treat native- like pronunciation as the only accepted norm when rating speaking performance. Hence, Assessors of

pronunciation should focus mainly on whether speech is communicatively effective and understandable (Isaacs & Trofimovich, 2012; Jenkins, 2000).

Vocabulary and grammar are important, but they should not constitute the primary basis of speaking assessment. Grammatical correctness means the accurate and appropriate use of language structures, whereas vocabulary use refers to lexical variety, exactness, relevance, and flexibility. Nevertheless, researchers argue that speaking assessment should not be limited to linguistic accuracy alone (Bachman, 1990; Fulcher, 2003). Missing some exact words or committing minor grammatical errors may not impede effective communication. For this reason, examiners should assess vocabulary and grammar as a part of a broader communicative ability that involves discourse management and interactional competence (Brown, 2004).

Furthermore, when rating monologic test tasks namely topic-based tests and oral presentations, examiners should take into account coherence and content organization. These criteria include the speakers' ability to produce well-organized, relevant, and meaningful spoken discourse (Hughes, 2003). By contrast, interactional competence is required when performing dialogic spoken tasks as it shows the test takers' ability to initiate turns, respond appropriately, negotiate meaning, and sustain communication (O'Sullivan, 2002). Since the examinee's performance can be affected by the examiner or interlocutor behavior, interaction should be taken into consideration during rating speaking test performance (Galaczi, 2014;Lazaraton, 1996).

2.3. Scoring Rubrics for Speaking Ability Measurement

Rating scales, also called scoring rubrics, provide examiners with a well-structured framework for the implementation of assessment criteria. Holistic rating scales assign only one global score, while analytic rating scales assess separate aspects of spoken performance. Holistic scoring is easy to implement and efficient, yet it may not reveal important differences among constituents of speaking ability. Analytic scoring, on the other hand, is transparent and informative because it enables raters to score each component of speaking ability separately (Luoma, 2004). According to a recent empirical study conducted by Khabbazbashi and Galaczi (2020), the researchers prove that the type of scoring model selected can deeply impact measurement results in speaking tests.

Similarly, Upshur and Turner (1995) contend that the choice of rating scales affects the rater perception, understanding, and evaluation of speaking performance. A poorly constructed rating scale may generate ambiguity, while a well-designed scale can contribute to more reliable scoring. Likewise, studies on proficiency descriptors underscore the significance of clearly specified and empirically supported descriptors (North, 2000). With reference to the Algerian EFL context, Sadoune (2020) stressed the importance of using scoring rubrics in rating speaking performance at Ouargla University. Despite this, the use of rating scales alone does not ensure the quality of rating; raters should have a good understanding of descriptors and how to apply them in a consistent manner across test sessions and candidates (Kim, 2015; Wigglesworth, 1993).

2.4. Reliability and Rater Variability

Intra-rater reliability refers to the stability of the same rater's scoring decisions across time, whereas Inter-rater reliability concerns the extent to which different raters provide consistent scores to the same performance. When measuring speaking ability, both types of reliability are central because raters may differ in their degree of leniency or severity, assessment focus, and the application of assessment criteria (Hughes, 2003). Generalizability studies have highlighted that test task and rater characteristics can lead to significant variability when scoring second language speaking tests (In'nami & Koizumi, 2016).

The issue of rater variability has been largely observed in speaking performance measurement practices. Wigglesworth (1993) highlights the necessity of analyzing rater bias in scoring speaking ability. Further, Eckes (2005), depending on the use of many-facet Rasch measurement, asserts that rater effects like severity can greatly affect writing and speaking test scores. Moreover, Kim's doctoral research as concerns the improvement of rating competence and

her studies on rater behavior also indicate that raters may refer to different aspects of speaking ability and develop their rating ability in different ways over time (Kim, 2011, 2015). Such findings support McNamara's (1996) view that raters cannot be neutral like scoring machines and therefore they should be treated as interpretive agents during the scoring process.

2.5. Validity as a Measurement Criterion

Validity is also of equal concern in measuring speaking performance. As argued by Messick (1989), validity is relevant to test content, score interpretation and use. In so far as testing speaking is concerned, to ensure construct validity, examiners should use test tasks and rating criteria that properly capture the construct of speaking ability. Considering speaking ability as multifaceted construct suggests that a valid speaking assessment should not be narrowly based on fluency, grammar, and pronunciation alone (Fulcher, 2003). Content validity is realized when test tasks and assessment criteria match the learning objectives and expected contexts of language use (Bachman & Palmer, 1996).

2.6.The Importance of Fairness and Rater Training in Testing Speaking

Fairness has a close connection with validity and reliability. Fair speaking tests should give candidates sufficient and equal opportunities to demonstrate their capacities and should minimize construct-irrelevant variances that cause score variability (Kunnan, 2004; McNamara & Roever, 2006). Several factors can threaten fairness including unclear criteria, topic familiarity and difficulty, anxiety, inconsistent scoring, and rater bias. For instance, topic familiarity and background knowledge may affect speaking test performance and weaken score validity by introducing construct-irrelevant variances if not adequately managed (Khabbazbashi, 2017). Other factors that have no relation with speaking ability can influence the judgement of raters: classroom participation, accent, confidence, and familiarity with the candidate (Derwing & Munro, 2005).

Rater training is widely recommended as a valuable solution for enhancing rating consistency and minimizing unwanted variability. Training of raters can involve familiarization with scales, discussing benchmark performances, practicing rating, ensuring calibration and moderation procedures (Fulcher, 2003; Luoma, 2004). In this context, researchers see that training can improve examiners' assessment literacy and reduce rater variability, although it cannot fully eliminate all sources of variation among raters (Lumley & McNamara, 1995; Weigle, 1998). Consequently, rater training should be reinforced with clear rating scales, moderation procedures and more than one rater if feasible especially in important speaking performance tests.

2.7.Challenges in Rating Speaking Tests

Previous research on speaking ability measurement has indicated that teachers frequently confront some practical challenges during testing speaking, such as restricted test time, large numbers of students, lack of training, unclear rubrics, and limited assessment literacy (Fulcher, 2010; Green, 2014). These constraints can hinder teachers' ability to implement standardized assessment procedure, despite their awareness of the value of reliability and fairness (Brown, 2004).Similar issues have been discussed in Algerian studies. First, as reported by Sadoune (2020), Rubrics are required for speaking performance assessment at the University of Ouargla. Second, Senouci (2024) explored teachers' practices in assessing speaking in Algerian university EFL classrooms and highlighted many weaknesses as reliance on oral presentations, imbalance in assessment criteria, and limited assessment literacy. Third, Bouab and Bentayeb (2023) found that the multi-componential nature of speaking, scoring inconsistencies, time restrictions, and unauthentic speaking tasks are serious challenges in speaking assessment.

3. Methodology

3.1. Study Design

This research employs a qualitative descriptive case study design. Depending on a qualitative approach is suitable for the study because our purpose is to investigate teachers' practices, perceptions and rating challenges in a natural university setting (Creswell, 2014; Dörnyei, 2007). The descriptive nature of the study is also appropriate since the researchers intend

to provide a comprehensive account of existing speaking test rating practices at the University of El Oued EFL department (Merriam & Tisdell, 2016).

Method triangulation was adopted in the study through the combination of non-participant observation with semi-structured interviews. In qualitative research, triangulation is commonly used to gain a broader understanding of a phenomenon by using multiple data methods or sources (Carter et al., 2014; Denzin, 1978; Noble & Heale, 2019). In this study, observations documented what actually occurred during speaking tests, while interviews elicited teachers' interpretations, perceptions and suggestions for improvement. Using both research tools enabled us to compare and interpret observed rating practices with teacher responses together.

3.2. Research Context and Participants

The study was carried out at the University of El Oued EFL department in Algeria at the end of the first semester of the academic year 2025/2026. The selection of this setting is because testing speaking forms an integral part of EFL education, where students are required to develop their speaking skills for presentations, academic participation, oral discussions, and prospective professional interaction. The study participants were 8 EFL teachers who had experience in rating speaking test performance at the university under investigation. They were 3 males and 5 females aged between 25 and 53 years old. Their experience with testing speaking ranges between 3 and 10 years. The teachers were selected purposively according to their direct experience with speaking test administration and rating. Purposive sampling is justified in qualitative inquiries when participants are chosen for their wide knowledge and relevance to the phenomenon being investigated and the research problem (Creswell, 2014; Patton, 2015).

3.3. Data Collection Instruments

3.3.1. Non-Participant Observation

The rationale for using non-participant observation is that it allows the researchers to document current speaking test practices without affecting the measurement process. Mackey and Gass (2016) claim that observation is a useful tool in qualitative studies as it helps the researcher in recording interactions, procedures and events as they occur naturally. The observation guide addressed speaking test format, types of tasks, topic selection, time allocation, teachers roles, scoring tools, use of rating scales, recording practices, presence or absence of moderation, and learner anxiety.

3.3.2. Semi-Structured Interviews

Semi-structured interviews were conducted with 8 EFL teachers to explore their perspectives on speaking test rating practices. The 12 interview questions were about rating criteria, scoring procedures, rubric use, rating challenges, topic effects, rater subjectivity, learner anxiety, reliability, rater variability, fairness, and suggestions for improvement. Using semi-structured interviews was appropriate for our study because they allowed us to guide the discussion during the interviews while providing participants with the opportunity to reflect and elaborate on their experiences and views (Dörnyei, 2007).

3.4. Data Collection Procedures

The process of data collection involved two phases. Firstly, the researchers observed speaking test sessions using the observation checklist to document field notes. Secondly, semi-structured interviews were held face-to-face with teachers and lasted between 15 and 20 minutes following the observation phase. The interviews were audio-recorded with the participants' approval for the sake of proper data documentation and transcription. Interviewing participants after observation enables the researchers to pose focused questions based on actual observed practices. In fact, the interviews complemented the observation findings by uncovering how teachers clarified and justified their assessment practices. In addition, the participants were informed about the study purpose and assured that identities would be kept confidential. Participation in the study was voluntary and research data were used solely for academic purposes.

3.5. Data Analysis Procedures

The collected data were subjected to thematic analysis. This latter is defined by Braun and Clarke (2006) as a method that depends on the identification, analysis, and reporting of emerging patters within the obtained research data. The researchers adopted a hybrid deductive-inductive approach to analyze data. The initial coding was based on the review of literature and research questions. Additional codes emerged from the observation records and interview data. Such a hybrid approach is in line with Fereday and Muir-Cochrane's (2006) model that supports the combination of theory-driven and data-driven coding of themes.

The analysis included various steps: data familiarization, initial code generation, clustering of relevant codes, the development of themes, reviewing themes, and interpreting them according to the existing literature. The data obtained from observation and interviews were first analyzed separately and then compared so as to determine the areas of divergence and convergence.

3.6. Research Trustworthiness and Ethical Considerations

Lincoln and Guba (1985) highlight that research trustworthiness is assured through four criteria: credibility, transferability, dependability, and confirmability. The researchers used method triangulation when combining data from interviews and observation in order to enhance the study credibility. A detailed description of the study setting, participants, tools, and procedures was provided for supporting transferability. Further, the use of systematic coding strategies strengthened the dependability of this study. Besides, the promotion of confirmability was through basing data interpretation on observed facts and teachers responses. Equally important, the researchers tried to respect the principles of ethicality throughout the study. Participants were made aware of the research purpose, their participation was voluntary, and their confidentiality was assured. No student or teacher names were mentioned during reporting of the study findings.

4. Results and Discussion

4.1. Summary of the Identified Themes

Five main themes emerged from the thematic analysis of observation and interview data:

1. *Variability in scoring practices and rating criteria.*
2. *Lack of unified departmental rating standards.*
3. *The teacher' dual role as interlocutor and rater.*
4. *Procedural issues affecting speaking test reliability.*
5. *Proposed measures for improving speaking test scoring.*

The five themes will be discussed below with regard to the research questions and available relevant literature.

4.2. Variability in Scoring Practices and Rating Criteria

The findings of the study reveal that teachers use different criteria when rating speaking tests. These involve grammatical correctness, pronunciation, fluency, vocabulary range, coherence, and effective communication. This reflects the multidimensional construct of speaking ability as advocated by researchers (Fulcher, 2003; Luoma, 2004). However, the use, understanding, application and importance of each assessment criteria vary from one teacher to another. More precisely, some teachers give special attention to communicative effectiveness and the aspect of fluency as one teacher advances: "I personally prioritize the criteria of fluency because at the end of the day we as teachers are expected to form a fluent student and I think this is the aim of teaching speaking to EFL learners". Other teachers focus more on pronunciation and grammar because such criteria are required in academic English in EFL classrooms. In this vein, during interviewing teachers, one of them declares: "for me, much value and attention should be given to grammar accuracy and pronunciation during rating speaking tests because both of them are taught at university. Also, most of the EFL department graduates will go to teaching professions where they need these two criteria". Based on the interview data, these differences among teachers stem from previous experience, module objectives, personal beliefs, and absence of departmental guidelines.

More importantly, all the interviewed teachers state that they do not refer to or use rating scales during speaking tests. Some teachers do not even understand the meaning of a rating scale or its components. Other teachers cannot differentiate between analytic and holistic rating. Just two teachers demonstrated their knowledge of rating scales, but they say we cannot use them in rating spoken performances of students because of time restrictions, complexity of the scales, large number of candidates, and the temporary nature of speaking performance. An interviewee says: "I cannot conduct analytic rating in five minutes, test time is limited and I need some kind of training on how to apply it appropriately". This was observed by researchers during speaking test rating sessions when teachers assign a total score after a candidate finishes his/her spoken performance turn. In reality, no holistic or analytic scales were used by teachers.

The absence of rating scales and variation in selecting scoring criteria among teachers affect the way scores are assigned and what raters notice when rating speaking test performance of students. Using analytic scoring allows raters to focus on separate aspects of oral performance, whereas, depending on holistic scoring encourages raters to rely more heavily on overall judgement (Brown, 2004). The documented differences among teachers' rating practices and criteria suggest that candidates' performance may be rated according to various operational understanding of speaking ability, which can undermine rating reliability, validity and fairness. This resonates with previous studies indicating that raters of oral ability may vary in the criteria they observe and in the importance they assign to them (Eckes, 2005; Kim, 2015; Wigglesworth, 1993).

4.3. Lack of Unified Departmental Rating Standards

The findings from observations and interviews showed no clear evidence of a unified departmental rubric or shared rating specifications. Rather, scoring seems to depend widely on personal teachers' criteria, guidelines, and checklists, rating schemes or general impression. One teacher states: "Our institution does not provide us with a standard scoring rubric and we have to choose or develop our own rating criteria". Similarly, teachers consider the absence of standardized rubrics as problematic and they see that unified departmental rating criteria would promote test reliability. This aligns with Luoma (2004) who confirms that the lack of standardized criteria for rating speaking ability can adversely influence the scoring process as criteria operationalize the construct of speaking being measured.

Moreover, the unavailability of common descriptors led teachers to interpret some terms such as pronunciation, fluency differently. For example, some teachers associate fluency with smoothness, while other teachers understand it as speed and confidence. These differences can compromise test reliability and fairness (Derwing & Munro, 2005, Kim, 2015). This finding is compatible with other results of Algerian researchers. Sadoune (2020) stressed the usefulness and importance of rubrics in assessing speaking performance at University, while Bouab and Bentayeb (2023) identified unclear assessment criteria as major problem in conducting summative speaking tests. The present research enriches this discussion by proving that the absence of standardized criteria while rating speaking tests is a visible feature in Algerian EFL context.

4.4. The Teachers' Dual Role as Interlocutor and Rater

A prominent finding from observations was that the teacher acted as both interlocutor and rater at the same time. During speaking test sessions, just one teacher was responsible for managing test topics, asking follow up questions, controlling time, listening to the candidate, and assigning scores. The teachers' dual role adds to the work and cognitive load placed on them and can negatively influence scoring consistency. Interestingly, one teacher explains: "It was difficult to ask questions and give scores simultaneously", another teacher adds: "Assuming two roles as a teacher and rater at the same time is a challenge; I think that the presence of a second rater improves test quality and fairness". All the interviewed teachers highlighted that the absence of a second rater while scoring speaking tests results from established departmental practice and institutional limitations.

These findings are consistent with studies on Oral proficiency interviews, which have demonstrated that interlocutors can impact test taker performance through questioning techniques, support, and the way interaction is managed (Lazaraton, 1996). When the same teacher serves both as interlocutor and rater, the rating process are susceptible to subjective decisions and inconsistency (O'Sullivan, 2002). Furthermore, some of the interviewed teachers reported that single rating and the absence of a second rater or moderation procedures can weaken the validity of scores, especially when no unified rubric is used. This aligns with McNamara's (1996) argument that raters should be considered as active participants when assessing speaking performance rather than neutral rating systems or software.

4.5. Procedural Issues affecting Speaking Test Reliability

The observation allowed the researchers to report some procedural constraints. The duration of each speaking test was around five minutes per candidate; many students were tested within a short period; test topics were chosen from folded papers, and students' performances were not recorded. All these practices may impact test validity and reliability.

Some teachers claimed that short speaking test time may not provide a good evidence for measuring speaking ability. This corresponds with Bachman and Palmer (1996) who view that a brief oral production may demonstrate some aspects as pronunciation or initial fluency; however, it cannot give sufficient evidence of interactional strategies or discourse development. Besides, one teacher states: "I am responsible for testing and rating around 35 students in 3 hours in one day. I am really exhausted". These findings support previous research showing that when a large number of candidates are tested in succession, rater fatigue and tiredness may adversely affect intra-rater reliability (Hughes, 2003; Weir, 2005).

The issue of topic selection from folded papers further poses fairness problems. Teachers see that random selection is fair, but research shows that topics can vary in familiarity, required vocabulary, and difficulty. According to Khabbazbashi (2017), test topics and background knowledge can impact speaking test production and may give rise to construct-irrelevant variance. In case of unbalanced test topics, student scores may be influenced by topic familiarity and thus the obtained scores do not reflect real speaking ability.

Teachers also consider the absence of recording as another limitation. One of the teachers said: "We cannot record students' oral performances due to some ethical and institutional procedures", while another declares: "recorded spoken performances allow us to check our rating again and again, but recording is prohibited in our faculty". Moreover, teachers add that immediate scoring of performances without recording them prevents teachers from reviewing uncertain cases, conducting moderation, or undertaking delayed rating. This finding is in parallel with other studies that support recording speaking production of candidates for more valid and reliable scoring (Fulcher, 2003; Luoma, 2004).

Test taker anxiety was also remarkable during speaking tests and confirmed through the interviews with teachers. Some students could not say one word as the test proceeds; others hesitated till the teacher intervened to help them through asking follow up questions. A teacher says: "Test anxiety is a devil that threatens students' performance during speaking tests". These data are compatible with previous studies that highlight the effects of anxiety on memory retrieval, fluency, grammar, and pronunciation, thereby influencing test performance regardless of actual speaking ability (Horwitz et al., 1986; MacIntyre & Gardner, 1994). Anxiety may result in low achievement that may not reflect candidates' speaking ability and therefore validity concerns arise.

4.6. Proposed Measures for Improving Speaking Test Scoring

In response to the reported and noticed challenges when rating speaking tests of students, the interviewed teachers provided some useful suggestion for improving speaking test scoring. The teachers explained that the following measures are necessary: developing context specific rubrics, providing explicit and clear test specifications, conducting rater training courses and seminars, allocating more time to speaking tests, integrating moderation procedures, ensuring balance in

topic difficulty, and strengthening institutional guidance and support. A teacher states: "to overcome such a difficulty in rating speaking skills, I recommend organizing rater training by experts and providing teachers of speaking with ready-made assessment rubrics. For me, Recording and moderation are ideal solutions, but they are not possible at our university now".

Rater training appears to be highly important as argued by other researchers who emphasize that training can enhance raters' knowledge and understanding of assessment criteria and decrease rating variability, though it may not reduce all sources of rater effects (Kim, 2011; Lumley & McNamara, 1995; Weigle, 1998). However, rater calibration should be accompanied with unified rubrics, benchmark oral performance, score review procedure, and periodic revision of scoring process. The study findings also highlight the necessity of aligning speaking module objectives, test tasks, and scoring criteria because without detailed and explicit test specifications, the scores obtained from speaking tests may depend on teachers' individual choices and interpretations. In this way, such results may not reflect a valid representation of the speaking construct. For this reason, strengthening the alignment of the aforementioned solutions would strengthen both score interpretation and content validity (Bachman & Palmer, 1996; Weir, 2005).

5. Conclusion

This research examined the practices, criteria and challenges relevant to speaking test performance rating at the University of El Oued EFL classrooms through the combination of two research instruments: Non-participant observation and semi-structured interviews. The study findings revealed that teachers generally depend on relevant speaking assessment criteria involving fluency, grammar, pronunciation, discourse organization, and communicative effectiveness. Nevertheless, the study indicates considerable differences in scoring practices, lack of using analytic or holistic scales or other kind of rubrics, absence of shared departmental assessment criteria, and total reliance of individual teacher judgement.

The study also uncovered various procedural and contextual challenges: teachers act as interlocutors to elicit speech from candidates and do the rating of speaking performances simultaneously, brief time is allotted to speaking test performance, variation is topic difficulty, numerous students are evaluated in compressed time duration, oral performance are not audio or video recorded, along with the absence of moderation procedures. These factors may compromise the validity and reliability of scores resulting from speaking tests.

On the whole, this investigation suggests that the measurement of speaking ability at the university level would improve if it shifts away from reliance on individualized rating practices toward depending on more standardized and institutionally supported assessment procedures. Well designed and clear rating scales, scorer training, moderated rating, balanced task design, tolerated recording and documentation of oral performances are essential if examiners would like to obtain more consistent, defensible, and transparent speaking test scores.

6. Recommendations

With reference to the study findings and relevant previous literature, some recommendations are suggested so as to improve the rating of speaking performance tests and mitigate its challenges in higher education EFL contexts:

- *EFL departments should develop or adopt a standardized rating scale or a common scoring rubric suited to the speaking course objectives.*
- *Rater calibration workshops should be held on regular basis to enable teachers to interpret assessment criteria consistently and harmonize their judgement.*
- *It is advisable to introduce scoring moderation sessions before and after formal speaking tests because such scores affect final grades and evaluation.*
- *It is necessary to review and balance speaking test topics as regards their familiarity, difficulty, and communicative requirements.*
- *It would be useful to record oral performances for the sake of reviewing, second scoring, and rater training.*

- *Sufficient time should be devoted to speaking test rating to minimize examiner fatigue and elicit adequate speaking performance samples.*
- *It is important to inform students of the rating criteria prior to the test so that assessment requirements are explicit and stress is reduced.*
- *Institutions should provide guidelines that explain and clarify the stages of developing speaking test (design stage, operationalization stage, and administration stage).*
- *Teachers should use stress reduction strategies in order to create a decent atmosphere to alleviate candidates' anxiety during speaking performance tests.*
- *The place of speaking assessment in the Algerian educational system should be reconsidered. Speaking skills of students should be assessed even before joining university EFL classrooms.*
- *Automated scoring systems, machine learning and latest advances in artificial intelligence chatbots should be exploited by teachers and students in measuring and rating speaking performance as such systems provide instant feedback and grades.*

7. Study Limitations

Some limitations should be acknowledged in the present study. First, conducting research in one university department restricts the generalizability of the findings. Second, the study analyzes EFL teachers' practices and perceptions, but it does not involve students' views. Third, the researchers did not examine score sheets or candidates' marks, which could have revealed more evidence as concerns rating variation among teachers. Fourth, this study was not a longitudinal investigation, as it explored rating speaking test practices within one semester rather than over various semesters or academic years. Finally, the number of interviewed teachers was limited, which is suitable for qualitative study; however, it is not sufficient for statistically generalizable claims.

8. Directions for Future Research

Future studies could expand the scope of this study by examining practices for scoring EFL learners' oral performance across a number of Algerian universities. Further investigations could gather and evaluate recorded oral test performances for the sake of calculating intra-rater reliability more directly. Other empirical studies could assess the influence of rater training on scoring reliability. Additionally, studies on rubric development could create and validate a specific speaking scale tailored for Algerian EFL university students. More importantly, future research could consider learners' views on test anxiety, fairness, assessment criteria, and preferred test task types.

9. References

1. Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
2. Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.
3. Bouab, E. b, & Benettayeb, A. (2023). Assessment of the speaking skill in Algeria: Facts and Challenges. *Journal of Legal and Social Studies*, 8 (1), 18-32.
4. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa
5. Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Pearson Education.
6. Bygate, M. (1987). *Speaking*. Oxford University Press.
7. Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2–27). Longman.

8. 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/I.1.1
9. Carter, N., Bryant-Lukosius, D., DiCenso, A., Blythe, J., & Neville, A. J. (2014). The use of triangulation in qualitative research. *Oncology Nursing Forum, 41*(5), 545–547. https://doi.org/10.1188/14.ONF.545-547
10. Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics, 6*(2), 5–35.
11. Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE.
12. Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods* (2nd ed.). McGraw-Hill.
13. Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379–397. https://doi.org/10.2307/3588486
14. Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford University Press.
15. Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2
16. Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods, 5*(1), 80–92. https://doi.org/10.1177/160940690600500107
17. Fulcher, G. (2003). *Testing second language speaking*. Pearson Education.
18. Fulcher, G. (2010). *Practical language testing*. Hodder Education.
19. Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.
20. Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics, 35*(5), 553–574. https://doi.org/10.1093/applin/amt017
21. Green, A. (2014). *Exploring language assessment and testing: Language in action*. Routledge.
22. Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70*(2), 125–132. https://doi.org/10.2307/327317
23. Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.
24. In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing, 33*(3), 341–366. https://doi.org/10.1177/0265532215587390
25. 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. https://doi.org/10.1017/S0272263112000150
26. 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. https://doi.org/10.1093/applin/amm017
27. Jenkins, J. (2000). *The phonology of English as an international language*. Oxford University Press.
28. Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing, 34*(1), 23–48. https://doi.org/10.1177/0265532215595666

29. Khabbazbashi, N., & Galaczi, E. D. (2020). A comparison of holistic, analytic, and part marking models in speaking assessment. *Language Testing, 37*(3), 333–360. https://doi.org/10.1177/0265532219898635
30. Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking assessment* [Doctoral dissertation, Teachers College, Columbia University]. ProQuest Dissertations and Theses.
31. Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly, 12*(3), 239–261. https://doi.org/10.1080/15434303.2015.1049353
32. Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. J. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge University Press.
33. Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing, 13*(2), 151–172. https://doi.org/10.1177/026553229601300202
34. Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE.
35. Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54–71. https://doi.org/10.1177/026553229501200104
36. Luoma, S. (2004). *Assessing speaking*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733017
37. Mackey, A., & Gass, S. M. (2016). *Second language research: Methodology and design* (2nd ed.). Routledge.
38. MacIntyre, P. D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning, 44*(2), 283–305. https://doi.org/10.1111/j.1467-1770.1994.tb01103.x
39. McNamara, T. F. (1996). *Measuring second language performance*. Longman.
40. McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Blackwell.
41. Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research: A guide to design and implementation* (4th ed.). Jossey-Bass.
42. Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education/Macmillan.
43. Noble, H., & Heale, R. (2019). Triangulation in research, with examples. *Evidence-Based Nursing, 22*(3), 67–68. https://doi.org/10.1136/ebnurs-2019-103145
44. North, B. (2000). *The development of a common framework scale of language proficiency*. Peter Lang.
45. O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277–295. https://doi.org/10.1191/0265532202lt205oa
46. Patton, M. Q. (2015). *Qualitative research & evaluation methods* (4th ed.). SAGE.
47. Sadoune, F. (2020). The use of rubrics in assessing students' oral performance: Case of oral expression classes of Kasdi Merbah University–Ouargla. *Journal of the Researcher in Humanities and Social Sciences, 12*(6), 799–806.
48. Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.
49. Senouci, M. (2024). The assessment of oral expression in the EFL classroom: Teachers' practices. *Aleph, 11*(1), 227–240.
50. Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.
51. Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). John Benjamins.
52. Taylor, L. (Ed.). (2011). *Examining speaking: Research and practice in assessing second language speaking*. Cambridge University Press.

53. Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3–12. https://doi.org/10.1093/elt/49.1.3
54. Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287. https://doi.org/10.1177/026553229801500205
55. Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.
56. Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10*(3), 305–335. https://doi.org/10.1177/026553229301000306