Journal for Educators, Teachers and Trainers
The PROFESIOLab electronic, peer-reviewed, open-access Magazine

ISSN 1989-9572
DOI:10.47750/jett.2024.15.03.022

Challenges Facing EFL Teachers in Conducting Summative Speaking Tests: Insights from the University of El Oued English Department

Aissa BERREGUI [1]
Mohammed NAOUA[2]

JournalforEducators, TeachersandTrainers,Vol.15(3)

https://jett.labosfor.com/

Date of reception: 18 Apr 2024

Date of revision: 21 Aug 2024

Date of acceptance: 20 Sep 2024

Aissa BERREGUI, Mohammed NAOUA,.(2024).Challenges Facing EFL Teachers in Conducting Summative Speaking Tests: Insights from the University of El Oued English Department. ***Journal for Educators, Teachers and Trainers*****, Vol. 15 (3). 254-**270

[1]Phd Student, Department of English Language, Laboratory of Pragmatics and Discourse Analysis, University of El Oued, Algeria.
[2]Professor, Department of English Language, Laboratory of Pragmatics and Discourse Analysis, University of El Oued, Algeria.

Journal for Educators, Teachers and Trainers
The LabOSfor electronic, peer-reviewed, open-access Magazine

Journal for Educators, Teachers and Trainers, Vol. 15(3)
ISSN1989 –9572
https://jett.labosfor.com/

Challenges Facing EFL Teachers in Conducting Summative Speaking Tests: Insights from the University of El Oued English Department

Aissa BERREGUI [1]
Mohammed NAOUA [2]
[1] Phd Student, Department of English Language, Laboratory of Pragmatics and Discourse Analysis, University of El Oued, Algeria.berregui-aissa@univ-eloued.dz
[2]Professor, Department of English Language, Laboratory of Pragmatics and Discourse Analysis, University of El Oued, Algeria.naoua-mohammed@univ-eloued.dz

ABSTRACT
The assessment of learners' speaking abilities is a difficult and complex undertaking for teachers in English as a Foreign Language (EFL) context like Algeria. The purpose of this study is twofold: First, to explore the primary challenges experienced by EFL teachers at the University of El Oued English Department in conducting summative tests of speaking. Second, to identify the potential solutions that can help in addressing the highlighted challenges and improving the quality of summative speaking assessment. To this end, the researchers opted for a qualitative approach depending on semi-structured interviews as tools for data collection. Through purposive sampling technique, ten EFL teachers were selected to participate in this study. Data were analyzed thematically so as to uncover recurrent themes and trends related to the research questions. The findings revealed that the multi-componential nature of speaking, ambiguity of assessment criteria, time limitations and subjective scoring were the most prominent problems. In addition, test anxiety, unauthentic tasks and scoring inconsistencies were also noted among the challenges that further complicate the assessment process. These problems can adversely undermine and impact the validity, reliability and practicality of summative speaking tests. In response, teachers suggested some solutions, such as the use of formative speaking assessment, provision of clear assessment criteria, rater training and technology integration to enhance the accuracy, objectivity and feasibility of speaking assessment. Moreover, this study underscores the need for reforms in designing speaking tests, institutional support, and adopting globally recognized assessment standards, along with leveraging artificial intelligence applications for more valid, consistent and fair summative speaking assessment. Future researchers should examine the effectiveness of technology-assisted speaking assessment in minimizing rater bias, ensuring scoring accuracy and enhancing assessment consistency.

Keywords: Challenges, EFL teachers, Practicality, Reliability, Summative Speaking Tests, Test validity.

1. Introduction

In tandem with the emergence of the communicative approach in the late 1970s, spoken language has gained much significance in language teaching. As a result, the assessment of speaking ability of second language

learners has become a cardinal issue in language testing practices (Fulcher, 2003; Nakamura, 1997).According to Hughes (2003), the assessment of oral abilities of English as Foreign Language (EFL) learners is imperative for providing a comprehensive evaluation of their overall oral proficiency and communicative competence. Summative speaking tests are developed for the sake of measuring learners' speaking skills after completion of a course or at the end of a teaching semester. This kind of tests is of paramount significance in the evaluation process (Brown, 2004). Nevertheless, the success and effectiveness of such tests may be undermined by multiple challenges and issues, such as test design, administration, rating procedures, and certain practical constraints (Kang, 2008; Underhill, 1989, Weir, 1990).

Summative testing of speaking, as stated by many applied linguists, encounters various obstacles that can undermine its valid and reliable assessment. Problems related to large classes and time constraints when administering a test, as well as to subjectivity in scoring are frequently reported by researchers (Heaton, 2003; Hussain et al., 2021; Madsen, 1983; McNamara, 1996). More importantly, teachers as assessors face difficulties in understanding the complex nature of spoken language aspects and choosing the right criteria for assessment, which can negatively impact the scoring process (Bachman & Palmer, 1996; Luoma, 2004). Additionally, other studies tackled problems in handling the complexity of speaking test tasks, which can weaken the quality of summative speaking tests. These aforementioned problems can result in inconsistencies in assigning test scores and difficulties in giving valid feedback to test takers (Alharbi&Surur, 2019; McNamara, 2000; Saefurrohman, 2018).

While the previously mentioned studies give us a general overview as regards the challenges and practices of implementing speaking tests, still there is a dearth of research tackling the assessment of speaking at the University of El Oued. Moreover, there is not enough investigation of workable solutions addressing the recognized challenges, especially at the level of Higher Education in Algeria. Available studies tend to discuss the obstacles without recommending effectives strategies and suggestions for enhancement. Therefore, the aim of the present research is to address this gap in the literature, not only by exploring the problems experienced by EFL instructors at El Oued University English Department in conducting summative speaking tests, but also by identifying practical measures proposed by the teachers to address the highlighted problems and improve the efficacy of summative speaking assessment. To this end, the researchers posed the following research questions:

1- What are the main challenges, which encounter EFL instructors at the University of El Oued English Department in conducting summative testing of speaking?
2- What possible solutions do these EFL teachers propose for overcoming the identified challenges and enhancing the quality of summative speaking assessment?

To answer these questions, the researchers formulated these hypotheses:

- ✓ EFL instructors at the University of El Oued English Department may face some challenges in testing their learners' oral proficiency.
- ✓ In order to address any problems undermining the quality of summative speaking tests, EFL teachers would recommend some practical solutions.

The incentive behind the choice of conducting the study at the University of El Oued was due to the fact that the English Department at the university under investigation has been implementing summative speaking tests for EFL learners for more than a decade. This period is considered enough to gain valuable insights into the process of testing speaking as well as the experience of the EFL instructors. At the research venue, testing speaking abilities of the students is done at the end of each academic term, which spans a period of three months for all levels. The oral proficiency test ranges from 5 to 10 minutes for each test taker.

By collecting useful qualitative data and insights from experienced EFL educators, this study seeks to enrich the overarching debate on upgrading the assessment of speaking in English and offer constructive recommendations for tackling this issue.

2. Literature Review

2.1. The Nature and Characteristics of Speaking Ability

Speaking is widely recognized as the most challenging skill to be mastered in any language owing to its complex and multifaceted nature. As claimed by Field (2011, p.70), it is the most "demanding of all human mental operations". Not only do speakers need to master language aspects like vocabulary and grammar, but they should also be aware of the sociolinguistic and discourse conventions required for speech comprehension and production. The difficulty is compounded by the different functions of speech, such as interactional and transactional talk. In addition, speaking is a social activity that represents an essential part of everyday life. Testing this skill is a difficult task as a result of the momentary nature of talk, the clarity of

speech sounds, the unique attributes of spoken grammar and lexis, along with the communicative and social aspects of talk (Luoma, 2004). Spoken language is dynamic, unpredictable and is also different from the written language in terms of structure, vocabulary, and discourse organization (Bygate, 2001; McCarthy &O'keefe, 2004).

Additionally, speaking is characterized by incomplete grammatical structures full of chunks, fillers, idea units with simple sentences compared to writing. Spoken English does not respect the logical word order due to the missing data that can be understood from the context of situation (Fulcher, 2003). Moreover, oral English includes hesitations markers, generic words, fixed phrases and tails (Carter & McCarthy, 1995; Chomsky, 1965). Speech is also characterized by numerous slips and errors including mispronunciation and incorrect choice of words owing to the lack of attention, which is commonly tolerated by first language speakers (Luoma, 2004). Generally, interactions are governed by social norms and contextual factors.

2.2. Qualities of a Useful Speaking Test

Language assessment is very necessary for measuring learners' language abilities and their learning progress. The effective assessment of language skills must be based on some principles to yield accurate, fair and useful results. The core principles that are fundamental in designing meaningful and workable language tests are validity, reliability, and practicality (Brown, 2004).

Validity entails the degree to which a test accurately measures the intended constructs, no more or less. A valid speaking test is expected to give accurate information regarding the candidates' oral communicative competence in real-life contexts. To be more precise, if the testers of oral proficiency use authentic test tasks, it will be easier for them to provide a precise interpretation of the test taker's speaking ability. For instance, oral proficiency interviews in face to face tests allow assessors to have a clear idea about the candidates' oral abilities in non-test situations. Despite this, live tests take longer time and are too difficult to administer in large classes with hundreds of candidates (Weir, 1994; Young, 2008).

Equally important is the principle of reliability, which refers to the "consistency in scores regardless of when and how many times a particular test is taken" (Li, 2011, p. 268). A reliable spoken test should produce similar results on two different times and its assessors are supposed to assign comparable scores to the same test over two distinct administrations. Therefore, to ensure a dependable test of speaking, assessment criteria should be in place coupled with the training of raters in using grading scales and various test formats in order to minimize scoring subjectivity and inconsistencies (Hughes, 2003; Runder, 2001).

Besides validity and reliability, the criterion of practicality in oral proficiency tests is another significant dimension. It stands for the feasibility of test administration and rating within the limited time frame and facilities (Chapelle& Douglass, 2006). Direct tests are considered less practical in achieving a valid speaking assessment. That is why the availability of resources is needed to enhance the practicality of direct speaking tests (Weir, 1994).

Overall, it can be concluded that the three principles are complementary. A valid test must be reliable, in that, increasing test reliability contributes to its validity (Li, 2011). Likewise, test validity can be compromised by practical issues. Gong (2010, p.7) posits that "a reliable and valid oral English test is also connected with how the speaking test is conducted". Nonetheless, ensuring higher levels of reliability and validity in oral tests may not be achievable due to practical considerations.

2.3. The Importance of Summative Assessment

Summative assessment, as per Brown (2004), is intended to measure students' performance following the finalization of a course or a teaching period. It includes formal tests for the sake of providing grades or accreditation. These tests gauge learners' achievement against pre-defined standards or criteria (Harlen, 2005). To illustrate, conducting a final test of speaking ability at the end of an academic term is a kind of summative assessment that evaluates the test takers' oral proficiency level and gives them grades (Hughes, 2003). Generally, summative tests, commonly referred to as assessment of learning, focus on measurement, grading and certification (Black & William, 2012).

Moreover, summative oral tests are very critical to the success of the teaching process. Their benefits are cited by Brown (2004) as follows:

- They ensure students accountability for their learning and meeting the educational goals.
- The implementation of summative speaking tests motivates learners to enhance their speaking skills so as to achieve good grades.
- Results of these tests can supply teachers with ample information regarding the effectiveness and quality of their instructional practices and educational materials.
- These tests can be used as benchmarks for measuring students' speaking abilities. Benchmarking

allows students to compare their performance against defined norms and helps in establishing fair and consistent assessment processes.

While summative tests of speaking play a vital role in the evaluation of overall language proficiency, they face various obstacles that can impact their efficacy and quality. These difficulties involve several issues pertaining to validity, reliability, assessment criteria, subjective scoring, rating scales and test administration constraints (Fulcher, 2003, Harlen, 2012)

2.4. Types of Speaking Tests

Students' oral proficiency is tested in three main ways: Direct, indirect or semi-direct methods. First, the indirect method was widely recognized before the inception of communicative language testing. This type of tests does not require students to talk. Instead, pronunciation tests were used to check the ability of examinees in identifying sounds or words that have different pronunciation (Lado, 1961). Second, in case of direct or live test methods, a test taker speaks with at least one interviewer or more interlocutors. A case in point is the oral proficiency interview (O'loughlin, 2001). Luoma (2004) argues that a live test of speaking is not highly practical as it consumes so much time to be conducted and it is not useful in demonstrating the candidates' true speaking abilities. To overcome these shortcomings, the use of pair interview is suggested to perform direct tests involving interaction between two examinees without any intervention from the examiner. Finally, the semi-direct testing format means that a large number of candidates are provided with the same instruction and prompt while undertaking the test via a computer, tape recorder or any other technological device in the absence of an interlocutor. Then, the examiner assigns scores to their performance. These kind of tests ensures fairness and practicality, but still unauthentic (Qian, 2009).

2.5. Challenges of Testing Speaking

The complexity and intricacy of speaking greatly influences the assessment of this skill. That is to say, the multi-componential nature of oral proficiency and its multiple aspects have to be taken into account in order to develop a valid and reliable speaking test (Taufiqulloh, 2009). As per Hughes (2003), the correct measurement of speaking ability is a challenging undertaking. In practice, it is the most difficult skill to test, in that it can cause various problems at all stages of test development. Obstacles may arise when selecting the element to be tested, choosing and preparing test tasks, determining assessment forms, and administering the speaking tests (Aleksandrzak, 2011).

2.5.1. Validity Issues

The question of validity is crucial when it comes to testing oral proficiency of EFL learners. Validity indicates that a test measures exactly what it purports to measure and nothing else (Hughes, 1990). Achieving a valid speaking test is a considerably intricate task because speaking includes the simultaneous engagement of a multitude of various abilities that may overlap and develop at varying rates (Harris, 1969). Accordingly, the design of speaking tests is a major challenge as it requires a careful selection of the information to be provided by both the tester and the testee, in addition to a good specification of the testing procedures and the test purpose (Norris, 2000). According to Luoma (2004), if a test is not valid it cannot provide reliable results. Cohen (1994) states that validity is problematic as learners' spoken discourse in the classroom or in test situations may not reflect their oral performance in real-world scenarios. In a similar vein, the difficulty of speaking assessment is due, in the main, to the "lack of understanding of what constitutes speaking ability" (Lado, 1961, p. 239). In essence, speaking is not well defined, the components to be tested are not clear. Referring to Fulcher (2003), the main problem in testing speaking is the issue of construct definition.

2.5.2. Construct –irrelevant Variances

Some factors may affect the test takers' spoken performance and the rating process during a test to a great extent. The type of interaction, methods of testing, choice of topic, the examiner effect, and characteristics of the candidates can lead to variability in scoring (Berry, 2007; Kunnan 1995; Shohamy, 1994). Similarly, test takers may suffer from second language anxiety which can negatively influence their oral performance during the test. This issue is proved to impact examiners' judgement as well as it is time consuming and frustrating (Woodrow, 2006). Hughes (1989) asserts that it is imperative to consider the psychological state of examinees and alleviate exam stress so that test takers can show their full potential and performance. As such, the provision of a decent and conducive atmosphere is required in testing speaking.

2.5.3. Topic Choice and Task Difficulty

The selection of the topic for each test task is of utmost importance because familiarity with the content may affect the candidates' output while performing the speaking test task. The examiner should choose topics that are appropriate to the students' level. That is to say, topics should correspond to the students' experiences, ages, cultures, cognitive abilities, and not too technical. Respecting these measures in topic selection will allow

test takers to gain better outcomes in their spoken output (Galaczi&ffrench, 2011). What is more, task difficulty is a matter of prime concern as far as testing oral proficiency is concerned. It can affect test validity and reliability to a great extent. Task complexity can falsely decrease or increase the learners' performance in speaking tests, thereby undermining assessment integrity because the attained results do not reflect the test taker's true speaking abilities (Luoma, 2004).

The main challenge related to task difficulty is designing or choosing test tasks that suit students' proficiency levels. Too difficult tasks can cause learners' anxiety or restrict their potential in demonstrating their oral skills. Contrariwise, too easy tasks may not accurately elicit the various speaking abilities of examinees (Skehan, 1998). Furthermore, the issue of task difficulty can also impact the validity of oral proficiency tests (Messick, 1996). Put another way, if the selected tasks do not simulate the language used in real-life contexts, the speaking test may fail to measure the targeted construct. Robinson (2001) asserts that cognitively demanding tasks can advantage some students at the expense of others leading to the distortion of the assessment results. Not only that, but due to the characteristics of oral tests, the question of task difficulty may influence the candidate's performance. Various tasks have the potential to generate diverse language aspects and modes of interaction, causing difficulties in achieving consistent results across different speaking test formats. This difference can contribute to test unreliability (Ellis, 2003).

In order to overcome these issues, meticulous task development and pilot testing are necessary. Careful needs analysis is required prior to task design to ascertain the alignment of tasks with the learners' level of proficiency and authentic language use in real-life situations. Besides, piloting sample test tasks can reveal possible problems associated with task difficulty and helps in implementing necessary changes before test administration. In this way, language testers can develop more fair and accurate tests of speaking that optimally measure learners' real language abilities (Fulcher, 2003).

2.5.4. The Issue of Reliability

Reliability means that the test yields stable scores over various administrations. It represents one of the great problems when conducting speaking tests. According to Ur (1996), reliability of oral tests is at stake since there are notable variations as regards scorers' judgements. The difference in scores can result from discrepancies between raters, differences in implementing the same test as well as insufficient guidelines or assessment criteria. Concerning this matter, Kuo and Jian (1997) mention that test taker-related reliability stands for the bad performance of the examinees due to some factors, such as personal issues or sickness. This performance does not reflect the true level of the students resulting in an unreliable assessment. Besides, another reason is rater-reliability which implies scoring errors committed by human raters during the oral test. This latter happens in interview tasks when the rater is also the interviewer who sets the appropriate level for the test taker while testing oral proficiency (Fulcher, 2010).

More importantly, during the test, speech is elicited by means of tasks in the presence of a human assessor who refers to a scoring rubric so as to give a score that reflects the test taker's ability. In this kind of performance assessment, errors in measurements are inevitable because of task variability or differences in rater decisions (Backman et al., 1995; Wigglesworth, 1993). Thus, test reliability can be improved by ensuring the examiners and test takers' familiarity with test methods, tasks and protocol. In addition, some measures should be applied including the evaluation of scorer reliability, the design of robust scoring rubrics, and providing training sessions for raters for the sake of standardizing the assessment procedure (O'Sullivan, 2000).

2.5.5. Subjective Scoring

Rating the speaking ability of test takers by human scorers raises the issue of subjectivity (Alderson, Clapham& Wall, 1995). For this reason, the selection of raters is decisive despite the fact that subjective judgements are not easily avoidable (Heaton, 1989). In this connection, Brown (1996) contends that subjective scoring can cause examiner disparities that do affect the scores and impact rater reliability in a negative way. Furthermore, the availability of more than one scorer can also influence test reliability. With reference to Underhill (1998), "the more assessors you have…the more reliable the score will be" (p.89). On top of that, when the rater plays the role of an interlocutor, it is harder for this rater to assign an accurate score to the examinees while communicating with them as well (Weir, 1994). Besides, it is not possible for the rater to have the same scoring performance and concentration from the beginning till the end of the speaking tests because oral assessment is time consuming, mainly with a high number of test takers (Çopur, 2002). These considerations should be taken into account in the process of designing and implementing tests of oral language ability.

2.5.6. Assessment Criteria

One of the major limitations in developing speaking tests is the establishment of the necessary assessment criteria. Due to the vague nature of speaking ability, selecting the constructs to be measured is still a big question. In other words, it is not clear whether testers should focus on fluency, accuracy or other language aspects when testing oral performance. Further, the value and grading of each component is a subject of disagreement among testers (Kiato&Kiato, 1996). The development of a rating sale, its proper application and adherence to it are other persistent challenges (Madsen, 1983).

2.5.7. Rating Scales

A rating scale refers to the concise description of each level of speaking ability like B1, B2, C1, and C2 in the Common European Framework of Reference for languages (CEFR). Scoring rubrics provide brief description of the expected performance of test takers at each level to assist the assessors in determining the level or score of the student in a test (Council of Europe, 2001; Harmer, 2004; Underhill, 1998). Holistic and analytic scales are the two prominent types of rubrics that are prevalently used in measuring spoken language ability. The former includes a general description of the overall oral ability; whereas the latter involves more details about the various dimensions and aspects of the speaking skill like the choice of vocabulary, grammatical structures, and pronunciation (Brookhart, 2013). Notwithstanding, achieving a highly reliable and valid system of scoring is a real challenge, especially in testing oral proficiency. Testers are confronted with some difficulties in developing assessment scales that can be implemented in an objective manner. As a result, innovative technology can be exploited in scoring speaking tests, but scoring inconsistencies are inevitable due to disagreement over the importance of each aspect of speaking at the level of the scale itself (Al-Amri, 2010).

2.5.8. Test Administration Issues

Researchers have reported some practical constraints that may cause difficulties when administering a test of speaking. These encompass the need for a sufficient number of testers to cover the rating of large number of examinees, time allocated for implementing the test, tools and resources required for testing and facilities necessary for rater training (Cohen, 1994; Weir, 1990). The testing environment is of great concern in oral testing because the test site can affect both the performance of test takers and the scoring procedures adversely. Hughes (1989) affirms that noise and interruption can have negative repercussions not only on candidates' performance, but also on assessors' decisions. Therefore, scoring oral tests should be conducted in a quiet and distraction-free place.

Additionally, time limitation is one of the biggest challenges in oral test administration compared to other paper and pencil tests. In speaking tests, each test taker is most of the time tested alone or ideally in pairs. In consequence, the process of test preparation should be done carefully and appropriately so that testing students would not take too much time (Güllüoğlu, 2004). In this regard, Field (2011) highlights that the time factor is crucial in the processing of information while producing speech. In speaking tests, the conceptualization and planning of utterances should be considered so as to obtain a well formed speech in terms of syntax and the selection of lexis. The planning time helps candidates in generating, expressing and organizing ideas, ultimately producing highly accurate and fluent speech.

2.6. The Integration of Technology in Testing Speaking

The incorporation of new technologies in speaking tests has improved the assessment process. Yet still it presents certain problems that have to be put into consideration.

On the one hand, technological integration in oral proficiency testing has some merits. To begin with, it enables the remote testing of speaking via online platforms that provide the opportunity for synchronous testing of real-time conversations. This can help institutions with a large number of students and offers flexibility for test takers and their examiners as well (Chapelle& Chung, 2010; Chapelle& Douglass, 2006). Further, innovative technologies like Artificial Intelligence (AI) and automated scoring mechanisms can serve in decreasing bias and subjective judgements in speaking assessment. These devices and applications utilize algorithms in assessing the different aspects of speech, such as fluency, accuracy and pronunciation to assure consistency and objectivity in performance measurement (Litman et al., 2018). This can contribute to the improvement of test reliability and guarantee uniform application of assessment criteria. Above all, digital tools can provide timely feedback and corrections that help students to identify their weaknesses and address them quickly to improve their speaking abilities (Wang et al., 2018).

On the other hand, despite the abovementioned advantages, technology integration is not without its shortcomings. Some challenges are echoed in the literature involving:

- Technical problems such as internet connectivity concerns can hinder the online testing of speaking

affecting the reliability of rating process during test administration. Hence, adequate resources and technical support should be provided to reduce interruptions and disturbances (Chapelle, 2001).

- Due to digital divide, some students may encounter some difficulties in accessing to the required technological tools. This can cause disparities in testing speaking (Selwyn, 2016). Therefore, equal access to technological resources is essential for ensuring fair assessment.
- Integrating technology in speaking tests raises the issue of confidentiality and test takers' privacy. The storage and transmission of test data digitally is prone to unauthorized access. Consequently, institutions have to apply strict security systems to preserve examinees' information (Litman et al., 2018).

Against this background, it can be stated that a technology-based speaking test is a two-edged sword. However, for its effective implementation, teachers and learners need to receive adequate training in using technological instruments and programs (Selwyn, 2016).

3. Methodology

3.1. Study Design

The researchers opted for a qualitative design to investigate the different challenges experienced by EFL teachers in conducting summative tests of speaking at El Oued University English Department and to explore the possible solutions suggested by these teachers to overcome the existing challenges and enhance the effectiveness of summative speaking assessment. The choice of the qualitative research approach is motivated by the fact that it is more appropriate for conducting this study as it provides a profound exploration of the viewpoints and experiences of EFL instructors concerning the problems encountered during all the stages of speaking tests and the recommended improvements or solutions to address these issues (Creswell & Poth, 2018).

3.2. Research Site and Participants

This study took place in the Department of English at the University of El Oued, Algeria during the academic years 2023-2024. A cohort of 10 full-time EFL instructors, responsible for teaching the different English courses across all levels, took part in this research. The participants speak Arabic as their mother tongue and are aged within the range of 27 and 48 years. All of them hold a Doctorate in English and their teaching experience in EFL classes spans from 5 to 20 years. The study participants were six female and four male English teachers who declared that their expertise in teaching and assessing speaking extends beyond three years. These teachers are familiar with the process of preparing, administering and measuring summative speaking tests. The researchers contacted the participants by phone and email to request their consent to participate voluntarily in the interview of the present study.

3.3. Data Collection Procedures

The research data were retrieved through semi-structured interviews. The motive behind choosing this type of interview format was its flexibility in delving into specific topics while ensuring consistency in the structure throughout all interviews (Kvale, 2007). All the interviews where held face-to-face and took between 25 and 30 minutes. Moreover, the participants were interviewed in English in the teachers' room within the Department of English at El Oued University. The interview questions were generated by the researchers based on the review of related literature to obtain in-depth responses regarding the challenges confronted in developing and conducting summative oral tests, the problems of existing assessment criteria and rating procedures, in addition to possible solutions to enhance the process of testing speaking. Further, all interviews were systematically audio-recorded with the teachers' approval to ensure rigorous data recording and transcription. Alongside this, the participants were asked to rank the challenges and solutions they highlighted during the interview process according to their order of significance and gravity so as to pinpoint the main themes and the subordinate ones.

3.4. Analysis Techniques

The recorded interview data underwent a process of verbatim transcription followed by a thematic analysis. As noted by (Braun & Clarke, 2006), this method is effective in identification, analysis, and subsequent reporting of themes and trends inferred from the available data. A qualitative and exploratory analysis was employed for the interview data. The analysis included three main phases: coding, theme development, and interpretation of the findings (Miles et al., 2014). Initially, the collected data were coded and categorized to generate important statements related to the research questions. Then, these codes were used in developing clearly defined themes to ensure agreement and consistency between the two researchers.

After that, the analysis process revealed important themes that may represent the three parts of the interview: Encountered challenges, problems of assessment criteria and rating scales, along with potential solutions respectively. Overall, the thematic analysis helped the researchers in detecting the recurrent

problems and possible solutions proposed by the participants. To enhance the reliability of data analysis, the two researchers not only cooperated in data coding, but also in double-checking the recognized themes and addressing any inconsistencies (Patton, 2015). At the final stage, the researchers computed and calculated frequencies and percentages to achieve a detailed quantification of prevalent themes, thereby guiding the overall analysis (Gibbs, 2007).

4. Results and Discussion

The analysis of the findings is organized under two sub-sections corresponding to the two research questions, involving the main problems encountered in conducting summative speaking tests and the possible solution proposed by EFL teachers to overcome the existing challenges and improve the overall quality of these tests.

4.1. The Main Problems Encountered in Conducting Summative Speaking Tests

This section delves into the primary challenges experienced by EFL teachers at the University of El Oued in conducting summative speaking tests in English Department. The results of these challenges are outlined in Table 1 below:

Table 1: The Main Problems Encountered in Conducting Summative Speaking Tests

Challenges	Frequencies	Percentages
The complex nature of speaking	10	100%
Unclear assessment criteria	9	90%
Time constraints	8	80%
Scoring subjectivity	8	80%
Test taker anxiety	7	70%
The use of non-authentic test tasks	7	70%
Inconsistent scoring	6	60%
The use of familiar topics and easy test tasks	5	50%

Table 1 provides a comprehensive summary of the eight identified challenges accompanied by the frequencies and percentages of EFL teachers who disclosed encountering these issues in conducting summative tests of speaking. Notably, the most prevalent challenge reported by all the interviewed teachers was related to the complex nature of speaking. Teachers stated issues pertaining to the multi-faceted nature of spoken language and the difficulty of choosing the right elements for testing speaking. They also indicated that speech is unpredictable and fleeting, and therefore assessors should be well concentrated and make great efforts to grasp the test taker's talk and assess it accurately. This finding corresponds with previous studies that highlight the fact that obstacles instructors face in designing reliable and valid speaking tests mainly stems from the multi-componential nature of this skill (Bygate, 2001; Field, 2011, Luoma, 2004).

Similarly, the ambiguity of assessment criteria was another remarkable problem echoed by 90% of the participants. This resonates with Bachman and Palmer (1996), who argue that the absence of clear and specific criteria for assessment may cause scoring inconsistencies, affecting both the validity and reliability of the speaking test. In this regard, teachers in this study confirmed that their institution does not supply them with clear guidelines about assessment criteria. Consequently, each teacher has or sets his own criteria to assess oral proficiency of his learners. As per these teachers, their criteria are not sufficient for capturing the full spectrum of speaking abilities, an issue reported by researchers like Hughes (2003), who asserts that assessment criteria must be well-defined, clear and in alignment with educational goals. Additionally, Brown (2004) claims that when there is no comprehensive criteria, assessors may fail to notice important constituents of speaking ability, resulting in incomplete assessment.

Furthermore, the problem of limited time was also prominent, with 80% of the participants highlighting the challenge of testing high number of students in a restricted period of time. This issue can lead to teachers' fatigue and lack of concentration. These factors can adversely affect teachers' rating performance and their accurate application of rating scales. These results align with the findings of Hussain et al. (2021), who note that test administration issues like time allotted for the test could influence the rater's decision. Besides, the teachers confirmed that the speaking test period for each student ranges from five to ten minutes. This

duration, as per the participants, is not sufficient for them to provide correct measurement of students' oral abilities because on the day of the test candidates may suffer from other socio-psychological problems, which hinder them to perform well. Teachers also mentioned that a student should be given adequate time to think and convey his message comfortably. This is in line with the studies of Güllüoğlu (2004) and Field (2011), who emphasize that time is an important factor in information processing during speaking tests.

In addition, the issue of scoring subjectivity was stressed by 80% of the teachers. This problem is due, in the main, to assessors' personal bias, which threatens objective assessment of test takers' speaking skills. Teachers highlighted that many factors can have an impact on their objective scoring of learners' oral abilities, such as students' body language, appearance, order in the test, familiarity with the examiner as well as teacher's mode during the test. The issue of subjective scoring is a well-documented challenge in speaking assessment where rating candidates' oral proficiency is done by human scorers (Alderson, Clapham& Wall, 1995). That is why; rater training and adherence to clear assessment criteria are required for improving the objectivity and reliability of the speaking test (Heaton, 1989).

Moreover, another notable challenge that was raised by more than half of the participants is related to test taker anxiety. This latter, according to the teachers, stems from the lack of formative assessment of learners' speaking abilities in class and the absence of pilot testing due to time constraints. Teachers stated that though they try to alleviate exam stress by being friendly, sympathetic, tolerant, encouraging and understanding before and during the test of speaking, but still students show high-levels of anxiety. This issue is considered as one of the construct-irrelevant variances that can undermine test validity and reliability as well (Fulcher, 2003; Kunnan, 1995). Test taker anxiety can negatively impact both the learners' performance during the speaking test and the assessors' rating accuracy to a great extent. This finding is emphasized by Woodrow (2006) and Hughes (1989), who claim that assessors should take into account psychological issues like examinee's test anxiety and put forward some strategies to reduce exam stress in order to allow students to give their best performance during speaking tests, and thus help examiners in the assessment process.

Equally important, the problem of using non-authentic test tasks was cited as a main concern by nearly 70% of teachers. Due to the short amount of time given for the speaking test (5 to 10 minutes), teachers commonly ask each test taker to choose one topic for discussion, prepare it and then talk about it in front of the teacher. This kind of assessment tasks lack authenticity, in that they cannot prepare learners for real-life speaking scenarios. That is, such tasks do not reflect the test taker's oral abilities in non-test situations. The issue of task authenticity was discussed by Cohen (1994), Hughes (1990), and O'loughlin (2001), who emphasize the need for authentic test tasks that mirror actual speaking situations outside the test context to ensure valid, useful and meaningful assessment for both students and teachers.

Above all, inconsistent scoring was identified by 60% of the participants as a crucial problem when testing spoken language proficiency. The teachers explained that inconsistencies in rating test takers' speaking abilities occur because there is no common unified rating scale. Some teachers rely on holistic scoring, whereas other teachers depend on analytic scoring. The problem is compounded by the difference in implementing these scoring rubrics which stems from the subjective interpretation of assessment criteria. This result is in accordance with Al-Amri (2010), who contends that scoring inconsistencies are widespread in speaking assessment due to the disagreement among raters over the significance of each constituent of speaking at the level of the rubric itself. This practice can undermine test reliability (McNamara, 2000). Therefore, ascertaining scoring consistency is necessary for achieving fair and accurate assessment, and can be enhanced through clear guidelines and good training of raters (Fulcher, 2010).

Last but not least, nearly half of the interviewed teachers declared that they usually use familiar topics and easy test tasks. The reason behind this is to avoid examinees' speaking block and reduce their test anxiety in order to make them feel comfortable and perform well during the speaking test. This practice diverges from Galacazi and ffrench (2001), who assert that the choice of topics for a speaking test should not be based on the test takers' familiarity with the topic, but it should correspond to their cognitive abilities and proficiency levels in order to allow assessors to provide accurate assessment of their students' speaking abilities.

Moreover, the use of easy test tasks in summative speaking assessment is problematic as it can lead to invalid and unreliable test results. In other words, using easy test tasks may enable examinees to do well and achieve good results that do not match their true speaking capacities. This also may cause inaccurate measurement of learners' oral proficiency and incomplete assessment (Luoma, 2004; Skehan, 1998). Further, the issue of easy test tasks can weaken the validity of speaking tests because such tasks may not allow for the correct measurement of the targeted spoken constructs (Messick, 1996).

4.2. Solutions Suggested by the Participants

The following section explores and discusses the solutions suggested by EFL instructors at El Oued University English Department to address the challenges they face in conducting summative tests of speaking and improve the quality of speaking assessment. These solutions are meticulously presented in Table 2 as follows:

Table 2: Solutions Suggested by the Participants

Solutions	Frequencies	Percentages
The use of formative speaking assessment	10	100 %
Development of clear assessment criteria	8	80 %
Training teachers on rating speaking tests	7	70 %
Integrating technology in speaking assessment	6	60 %

Table 2 summarizes the proposed solutions along with their frequencies and percentages in the teachers' responses. The analysis of the results reveals that teachers suggested four solutions for overcoming the highlighted problems and enhancing the effectiveness of summative speaking assessment.

Remarkably, there is a common consensus among teachers on the need for implementing formative assessment during speaking class time in order to prepare learners for summative speaking tests. Teachers believe that ongoing assessment should complement the summative one because the former allows students to get detailed and continuous feedback, which helps them to identify their strengths and weaknesses prior to taking summative tests. This finding aligns with previous researches that underscore the vital role of formative assessment in tracking and developing students' speaking skills (Brown, 2004; Harlen, 2012; Tang, 2016). The integration of formative assessment tasks enables teachers to measure their learners' progress and assists in alleviating their stress during summative tests. Consequently, students can show their full potential and true speaking abilities allowing teachers to provide accurate assessment. This can enhance test reliability, validity and fairness.

Additionally, provision of clear assessment criteria was another key solution proposed by 80% of the study participants. Teachers stressed that the lack of unified and transparent criteria poses difficulties in ensuring consistent and fair scoring. The availability of clear and well-defined assessment criteria allows students to understand what is required from them during the test and also helps teachers in designing speaking test tasks and measuring their students' performance. Studies by Bachman and Palmer (1996) and Hughes (2003) highlight the importance of well-structured criteria for enhancing the reliability and validity of oral proficiency assessment. In this vein, Brown (2004) and Fulcher (2003) assert that providing specific criteria that are aligned with the speaking constituents to be measured is essential for improving the accuracy and objectivity of the assessment process.

Moreover, seventy percent of the teachers suggested rater training as an effective measure for achieving accurate and fair scoring of speaking tests. Teachers stated that inadequate training may lead to inconsistent scoring, subjectivity and rater bias. This view is compatible with previous research which indicates that rater training helps teachers in understanding the required assessment criteria and the best scoring practices, thereby enhancing test reliability (O'sullivan, 2000; Wigglesworth, 1993). According to Bachman et al (1995), training programs help in ensuring that assessment criteria are applied uniformly across all raters. This contributes in reducing variability in rater's judgments, minimizing bias, and ensuring assessment credibility.

Finally, it goes without saying that more than half of the participants recommended the integration of some

technological tools like audio and video recordings in summative speaking assessment. Teachers see that recordings can facilitate accurate and objective measurement of students' speaking abilities. Put differently, such tools allow the teacher to record, analyze and review test taker's performance many times avoiding problems like fatigue or students' anxiety during live speaking tests. This finding supports recent studies that advocate technology-assisted speaking assessment for improving the practicality of administering oral tests and their reliable scoring (Litman et al., 2018; Selwyn, 2016; Wang et al, 2018). More importantly, teachers reported that the effective incorporation of technology should be accompanied with good planning and sufficient training in order to enhance the assessment of speaking.

5. Conclusion

The present study aimed at exploring the challenges faced by EFL teachers in conducting summative speaking tests at the University of El Oued English Department. The focus was on understanding the impact of the highlighted challenges on the reliability, validity and practicality of testing students' speaking abilities. Additionally, the researchers attempted to investigate the strategies and solutions proposed by teachers in order to address these problems and enhance the quality of speaking assessment. The findings revealed various significant issues and provide valuable suggestions for improving the overall quality of summative oral proficiency assessments.

The results of this study highlighted that EFL teachers encounter several challenges during the summative assessment of their learners' speaking skills. The most prevalent problems were the complex nature of speaking, ambiguous assessment criteria, time constraints and subjective scoring. Other notable challenges that further complicate the assessment process such as test taker anxiety, the use of unauthentic test tasks, inconsistent scoring, along with the use of familiar topics and easy test tasks were also stressed. These underlying issues can negatively impact the accuracy, objectivity, and feasibility of summative spoken language assessment. As for solutions, teachers regarded the implementation of formative assessment as the most successful strategy, followed by the development of well-defined assessment criteria, training of teachers on rating speaking tests and integrating technology in speaking assessment. The aforementioned findings accentuate the difficulty of testing speaking and highlight the need for workable solutions to improve the quality of summative speaking tests.

Though the current study offers valuable insights, some limitations need to be acknowledged. Firstly, the results of this research cannot be generalized to other contexts because the small sample size including ten EFL teachers from one university which has its specific testing procedure, resources and conditions. Therefore, studies in other educational institutions with different characteristics may yield alternative findings. Secondly, this study relied on self-reported qualitative data from semi-structured interviews. While these data were rich and detailed, but they can be susceptible to bias and may not completely represent the full range of nuances about participants' experiences and capture the difficulties of assessing speaking. Finally, the absence of a longitudinal approach limits our understanding of the evolution of such problems over time as changes occur in assessment procedures and educational strategies.

6. Recommendations

Some pedagogical implications and recommendations can be made from this study in order to overcome the challenges of testing speaking and enhance the summative assessment of EFL learners' speaking abilities at University level:

- Incorporating formative assessment of speaking skills throughout the academic semesters so as to track learners' progress, give them constant feedback and alleviate anxiety associated with high-stakes summative tests.
- Establishing clear assessment rubrics outlining detailed criteria for testing speaking to reduce confusion and scoring subjectivity. This can help in ensuring that all test takers are assessed on equal-footing using transparent and standardized criteria.
- The organization of regular training sessions or workshops for teachers on the accurate application of rating scales to mitigate scoring inaccuracies and bias as well as enhance inter-rater reliability in speaking

tests.

- Encouraging teachers to utilize audio or video recording devices to increase scoring reliability and ensure assessment fairness as these technological tools enable raters to review students' performances several times.
- The allocation of sufficient time for each speaking test so that examiners can provide accurate and complete assessment.
- Increasing test validity by designing authentic test tasks that reflect real-life speaking situations in order to motivate students and assess their true communicative abilities outside the test context.
- It is essential for teachers to use stress-reduction techniques to create a supportive and decent environment that helps in alleviating students' pressure during speaking tests.
- Encouraging alternative assessment of speaking, such as online tests that allow for flexibility and minimize the constraints of in-person speaking tests of large number of candidates.
- Teachers should upgrade their speaking assessment literacy through attending workshops and conferences to keep up with the latest innovations in speaking assessment.
- Institutions are required to align the criteria for assessing speaking with globally recognized benchmarks like the Common European Framework of Reference for languages (CEFR) to assure rigorous and internationally relevant speaking tests.
- Policy makers and test developers are recommended to reconsider the status of speaking in the Algerian educational system. Speaking should be assessed during early phases of education before students begin their university education.
- Teachers should explore the implementation of automated scoring systems, which use artificial intelligence and machine learning in assessing students' oral proficiency. Such systems can decrease teachers' workload, ensure consistent scoring, and offer timely feedback.
- Teachers are advised to ensure the alignment of speaking test tasks with the criteria of assessment in order to achieve accurate measurement of the targeted components. This sort of alignment will result in improved validity of oral assessment and give a clear representation of students' oral communicative competence.
- Students should exploit artificial intelligence chatbots and applications like Replika and Duolingo to practice and assess their speaking abilities prior to taking formal speaking tests.

7. Future Research

Further research can replicate the present study by investigating a larger sample of EFL speaking teachers from various Algerian or worldwide universities. Additionally, researchers are encouraged to examine students' perceptions and the difficulties they encounter in summative speaking tests. Furthermore, future studies may explore the effectiveness of rater training or integrating technologies like artificial intelligence in enhancing the validity, reliability, and practicality of speaking tests. Besides, longitudinal studies might be conducted to achieve valuable insights concerning the long-term effects of various speaking assessment practices and problems. Moreover, experimental studies can analyze the impact of speaking test tasks on students' oral performance. Similarly, researchers are required to conduct comparative studies between online and face to face oral proficiency tests in terms of test takers' performance and scores. Last but not least, further research is also needed to compare the procedures of summative testing of speaking in Algerian Universities against the standards of international English proficiency tests, such as IELTS or TOEFL. This could help in providing a benchmark for ameliorating existing summative testing frameworks.

References

Al-Amri, M. (2010). Direct Spoken English Testing is Still a Real Challenge to be Worthbothering About. *English Language Teaching, 3* (1), 113-117.

Alderson, J. C., Claphamn, C., & Wall, D. (1995).*Language testconstruction and evaluation*.Cambridge: Cambridge UniversityPress.

Aleksandrzak, M. (2011). Problems and challenges in teaching and learning speaking atadvanced level. *Glottodidactica, 37*, 37-48.

Alharbi, A. F., &Surur, R. S. (2019).The Effectiveness of Oral Assessment TechniquesUsed in EFL Classrooms in Saudi Arabia from Students and Teachers Point of View.English Language Teaching, 12 (5), 1-19.

Bachman, L., Lynch, B., & Mason, M. (1995). Investigating variability in tasks and raterjudgments in a performance test of foreign language speaking. *Language Testing, 12*, 238–57.http://dx.doi.org/10.1177/026553229501200206

Bachman, L. F. & Palmer, A. S. (1996).*Language testing in practice*. Oxford: Oxford University Press.

Berry, V. (2007). *Personality differences and oral test performance*. Frankfurt: Peter Lang.

Black, P., &Wiliam, D. (2012).The Reliability of Assessments. In J. Gardner (Ed.), *Assessment and Learning* (pp. 243-263). London: Sage.

Braun, V., & Clarke, V. (2006).Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Brown, J. D. (1996). *Testing in Language Programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York: Pearson Education, Inc.

Bygate, M. (2001).Speaking.In Carter, R. &Nunan, D. (2001).*The Cambridge Guide to Teaching English to Speakers of Other Languages*. Cambridge: CUP.

Carter, R. & McCarthy, M. (1995).Grammar of the Spoken Language. Applied Linguistics, 16 (2), 141-155.

Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing, and research*. Cambridge University Press.

Chapelle, C. A., & Douglas, D. (2006).*Assessing Language through Computer Technology*.Cambridge University Press.

Chapelle, C A and Chung, Y-R (2010) The promise of NLP and speech processing technologies in language assessment, *Language Testing* 27, 301–315.

Chomsky, N. 1965.*Aspects of the Theory of Syntax.*Cambridge, Mass: MIT Press.

Cohen, A. D. (1994). *Assessing language ability in the classroom.*Boston: Heinle&HeinlePublishers.

Council of Europe. (2001).*Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Çopur, D. (2002). Testing first year FLE students' oral performance using four speaking testmethods in spoken English II course and students' attitudes towards these speaking testing methods (Unpublished MA thesis).Middle East Technical University, Ankara.

Creswell, J. W., &Poth, C. N. (2018).*Qualitative inquiry and research design: Choosing among five approaches*(4th ed.). Thousand Oaks, CA: SAGE Publications.

Ellis, R. (2003). *Task-based language learning and teaching*.Oxford University Press.

Field, J (2011) Cognitive validity, in Taylor, L (Ed) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge UCLES/Cambridge University Press, 65–111.

Fulcher, G. (2003). *Testing second language speaking.*London:Pearson Longman Education.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

Galaczi, E., &ffrench, A. (2011).Context validity.In M. Milanovic& C. Weir (Series Eds) Linda Taylor (Vol. Ed).*Studies in Language Testing: Vol. 30. Examining speaking: Research and practice inassessing second language speaking.*Cambridge: Cambridge University Press.

Gibbs, G. R. (2007). *Analyzing qualitative data***.** London: SAGE Publications.

Gong, B. (2010, August 22–27). *Considerations of conducting spoken English tests for advanced college students*. Paper presented at the 36th International Association for Educational Assessment Conference, Bangkok. http://selectscore.com/fullpaperII19.pdf

Güllüoğlu, O. (2004). Attitudes and perceptions of the students at GaziUniversitytowards testing speaking (Unpublished MA thesis).Gazi University, Ankara.

Harlen, W. (2005). Teachers' summative practices and assessment for learning: Tensions and Synergies. *The Curriculum Journal, 16*(2), 207–223. https://doi.org/10.1080/09585170500136093

Harlen, W. (2012).On the relationship between assessment for formative and summativepurposes.In J. Gardner (Ed.), *Assessment and learning* (2nd ed., pp. 87-101). London:Sage.

Harmer, J. (2004).*The practice of English language teaching*.Longman.

Harris, D. P. (1969).*Testing English as a second language:* NewYork: Mc Graw-Hill Book Company.

Heaton, J. B. (1989). *Classroom Testing.*London; New York: Longman.

Heaton, J.B. (2003). *Writing English language tests*. London: Longman.

Hughes, A. (1989). *Testing for language teachers*.Cambridge:Cambridge University Press.

Hughes, A. (1990). *Testing for language teachers*.Glasgow:Cambridge University Press.

Hughes, A. (2003). *Testing for language teachers*.Cambridge University Press.

Hussain, S. Q., et al. (2021). Assessment of Oral Communication Skills of Students atTertiary Level by University Teachers in Pakistan.*İlköğretim Online- Elementary EducationOnline*, 20 (6), 668-684.

Kang, O. (2008). Ratings of L2 Oral Performance in English: Relative Impact of Rater characteristics and Acoustic Measures of Accentedness. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 6, 181–205.

Kitao, S.K. & Kitao, K. (1996).Testing speaking.*ERIC Document Reproduction,* ServiceNo. ED 398261, 1-7.

Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural modeling study* (Vol. 2).Cambridge: Cambridge University Press.

Kuo, J. & Jiang, X. (1997).Assessing the assessments: The OPI and the SOPI.*Foreign Language Annals*, *30*(4), 503-512.

Kvale, S. (2007).*Doing Interviews.*London: Sage Publications Ltd.

Lado, R. (1961). Language Testing: *The Construction and Use of Foreign Language Tests: A Teacher's Book.* New York: McGrow-HillBook Company.

Li, W. (2011). Validity Considerations in Designing an Oral Test. *Journal of Language Teaching and Research*, Vol. 2, No. 1, pp. 267-269, January 2011 © 2011 ACADEMY PUBLISHER Manufactured in Finland.

Litman, D., Strik, H., & Lim, G. S. (2018). Speech Technologies and the Assessment of Second Language Speaking: Approaches, Challenges, and Opportunities. *Language Assessment Quarterly*, *15*(3), 294–309. https://doi.org/10.1080/15434303.2018.1472265

Luoma, S. (2004).*Assessing speaking.*Cambridge: CambridgeUniversity Press.

Madsen, H.S. (1983). *Techniques in testing.*Oxford: OxfordUniversity Press.

Miles, M. B., Huberman, A. M., &Saldaña, J. (2014).*Qualitative data analysis: A methodsSourcebook*(3rd ed.). Thousand Oaks, CA: SAGE Publications.

Nakamura, Yuji. (1997). Establishing construct validity of an Englishspeaking test. *Journal of Communication, n6, 13-30.*

McCarthy, M. & O'Keeffe.A. (2004).Research in Teaching of Speaking.*Annual Review of Applied Linguistics*, 24, 26-43.

McNamara, T. (1996).*Measuring second language performance*. Chicago: AddisonWesley Longman.

McNamara, T. (2000).*Language testing.*Oxford: Oxford University Press.

Messick, S. (1996).Validity and washback in language testing.*Language Testing*, 13, 241-256.

Norris, J.M. (2000). Purposeful language assessment: Selecting the right alternative test. *English Teaching Forum, 38* (1), 18-22.

O'Loughlin, K. (2001). *The Equivalence of Direct and Semi-direct Speaking Tests*. Cambridge: CUP.

O'Sullivan, B. (2000). Towards a model of performance in oral language testing.(Unpublished doctoral dissertation).University of Reading, Berkshire. England.

Patton, M. Q. (2015). *Qualitative research and evaluation methods* (4th ed.). Thousand Oaks:Sage.

Qian, D. D. (2009). Comparing Direct and Semi-Direct Modes for Speaking Assessment: Affective Effects on Test Takers. *Language Assessment Quarterly*, *6*(2), 113–125. https://doi.org/10.1080/15434300902800059

Richards, J. C. (2008).*Teaching listening and speaking: From theory to practice*. Cambridge University Press.

Robinson, P (2001) Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA, in Robinson, P (Ed) *Cognition and Second Language Instruction*, Cambridge: Cambridge University Press, 287–318.

Rudner, L. (2001). *Reliability.*College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Saefurrohman. (2018). EFL Teachers Assessment Methods in Oral Communications.*Advances in Social Sciences, Education & Humanities Research (ASSEHR),* 27, 268-272.

Selwyn, N. (2016). *Education and technology: Key issues and debates* (2nd ed.). London: Bloomsbury Academic.

Shohamy, E. (1994). The Validity of Direct versus Semi-Direct Oral Tests.*Language Testing, 11*(2), 99-123.

Skehan, P. (1998). Processing perspectives on testing. In: *An Examination of Comprehensibility in a High Stakes Oral Proficiency Assessment for ProspectiveInternational Teaching Assistants* (p19*)*. Dissertation, McGregor L. A. The University of Texas, Austin.

Tang, L. (2016). Formative Assessment in Oral English Classroom and Alleviation of Speaking Apprehension. Theory and Practice in Language Studies, 6(4), 751. https://doi.org/10.17507/tpls.0604.12

Taufiqulloh, S. D. (2009). Designing speaking test.*Ekspianasi*, *4*, 183.

Underhill N. (1998). *Testing Spoken Language*. Cambridge: CUP.

Ur, P. (1996). *A Course in language teaching.*Cambridge: CambridgeUniversity Press.

Wang, Z, Zechner, K and Sun, Y (2018) Monitoring the performance of human and automated scores for spoken responses, *Language Testing* 35, 101–120.

Weir, C.J. (1990).*Communicative language testing*. New York:Prentice Hall.

Weir, C. (1994).*Understanding and Developing Language Tests.*London: Prentice Hall.

Wigglesworth, G. 1993. Exploring bias analysis as a tool for improving rater consistency inassessing oral interaction.*Language Testing* 10,3:305-35.

Woodrow, L. (2006). Anxiety and speaking English as a second language.*RELC Journal, 37*(3), 308-328.

Young, J. W. (2008). *Ensuring valid test content tests for English language learners*.R&D Connections 8. Princeton, NJ:Educational Testing Service.