Hindawi
Education Research International
Volume 2022, Article ID 8984330, 10 pages
https://doi.org/10.1155/2022/8984330

Research Article

The Reliability Analysis of Speaking Test in Computer-Assisted Language Learning (CALL) Environment

Raja Muhammad Ishtiaq Khan,[1] **Tribhuwan Kumar**,[2] **Ahmed Benyo**,[2] **Syed Farhat Jahara**,[3] **and Mir Mohammad Farooq Haidari**[4]

[1]*Majma'ah University, Zulfi, Saudi Arabia*
[2]*College of Science and Humanities at Sulail, Prince Sattam Bin Abdulaziz University, Al Kharj, Saudi Arabia*
[3]*Department of English Language and Translation, College of Sciences and Arts at Al-Asyah, Qassim University, Saudi Arabia*
[4]*Economic Faculty and Research Deputy of Taj University, Afghanistan*

Correspondence should be addressed to Tribhuwan Kumar; t.kumar@psau.edu.sa
and Mir Mohammad Farooq Haidari; mir.m.farooqhaidari@gmail.com

Received 9 February 2022; Revised 27 February 2022; Accepted 2 March 2022; Published 15 March 2022

Academic Editor: Ehsan Rezvani

Speaking ability is regarded as among the essential aspects of language development. Oral examination appeared challenging to evaluate due to the presence of human evaluators. The speaking method depends on the test's reliability, determined by the raters' scores. The current study is aimed at evaluating the speaking test's interrater reliability utilized to measure the speaking performance of Common First Year (CFY) students during remote learning. The data were obtained from 56 EFL learners using a scoring sheet and rubrics. Eight raters were responsible for rating the study. The speaking test's reliability was estimated using quantitative data analysis. Correlation coefficients and the Bland-Altman test were employed to assess raters' agreement. SPSS was utilized to analyze the data in this investigation. The study's findings suggested that the speaking exam used in the CFY program during remote learning has shown some reliability on correlations and acceptable norms on the Bland-Altman test.

1. Introduction

Education has changed rapidly in recent years, with a noticeable increase in e-learning, which involves teaching distantly or remotely and via online platforms [1–3]. Following the closure of all educational facilities in Saudi Arabia due to the COVID-19 breakout, an unforeseen quick transition from the typical "traditional" learning method [4, 5] to the newly e-learning supported, namely, online learning, has happened. Adedoyin and Soykan [6] argue that "without the pandemic's emergence, our education institutions would not have been so adept at online instruction." This suspension has also impacted the testing and evaluation system, partially testing language skills. Speaking is an essential skill in learning any language. It involves interaction and group discussion, which was significantly impacted by the outbreak of COVID-19. Institutions were bound to have an online evaluation during this era.

Language proficiency fosters collaboration and interaction among individuals from varied cultural origins in all spheres of life, education, and business in the globalized twenty-first century Kukulska-Hulme et al. [7]. Learning languages must consequently be a long-term commitment, implemented in a variety of ways to fulfill societal, professional, and educational objectives, as well as individual desires and needs. English is commonly recognized as the globe's lingua franca and the most widely spoken language [8–10]. Due to the importance and demand for the English language in the modern global era, English as a second language (ESL) students travel from all over the world to study the language [11, 12]. Consequently, significant effort has been expended in developing effective methods for learning

English. It will require a concerted, substantial, and remarkable effort on the part of both students and instructors [13]. The growth of communicative pedagogy has resulted in a greater emphasis on the fundamental goal of speaking proficiency in online learning. As a result, primarily speaking competence, testing speaking skills has been a prominent issue in the language development testing process [5, 14, 15]. Numerous limits apply, considering the nature of verbal skills. The essential issue in speaking testing ability is the necessity to detail the activities that construct a sample of the population of speaking activities while also demonstrating how the results of those tasks truly reflect the examinees' speaking capabilities [16–18]. Similarly, various factors affect our perception of how well someone can communicate vocally. Since the essence of verbal ability is still in its development, it is impossible to define it accurately. There is a contradiction that allows for assessing the many components of oral skill. Typically, speaking measures are based on pronunciation, vocabulary utilization, and correct use of grammar. Likewise, the speaking test will also assess relevancy and fluency [19]. Due to the variety of elements included in the speaking evaluation, its accurate judgment is not as straightforward as it is for the other skills [20]. Kang et al. [21] contend that there could be differences in evaluating oral skills because the test taker is required to utilize language in any way due to its interactive aspect. Additionally, because human examiners primarily conduct the speaking assessment, the speaking test's scoring is significantly biased. Kang and Kermad [22] highlight this point as the primary concern in speaking evaluation, as the subjective nature of the calculation of the scoring process may result in rater inconsistencies or shifts, affecting the test taker's scores and, conversely, rater reliability. Therefore, the grading criteria are vital for the speaking test [23, 24]. Speaking assessment also has certain practical constraints, contributing to the inconsistency of the outcomes, particularly in online learning. This includes time, a large number of examinees at the same period, operational costs, the attitude of the raters, the rater's training, the duration of the test, the usage of rubrics, and the average test's length [25]. Regardless of these constraints, school systems, universities, colleges, and standardized English agencies now assess examinees' oral proficiency. The oral performance is evaluated through various tasks involving presentations, group conversations, and role-playing, which are expected to elicit evidence regarding the evaluators' ability to communicate effectively [26].

The assessment process is highly dependent on a variety of factors that can hinder a learner's speaking skills during remote testing in situations like COVID-19, where rater and students have to engage through an online platform. Hence, the requirement of a well-defined, well-researched, and well-documented description of the exam results' trustworthiness is derived from logical, empirical evidence [27]. The language assessment process may also be centered on the correctness of the evaluations of learners' replies that may be supported by the premises of measure [28–30]. The purpose of language assessment is not merely to provide rating scales for awarding certain marks or levels of language ability but to explain the types of evidence that can be offered to justify the precision of the proficiencies of the grades [31]. Therefore, a speaking test procedure should be backed up by evidence that the test is performing the intended purpose. This entails presenting data on remote steps in addition to various reliability measures. Nonetheless, the research reveals that only a small portion of the validity question is addressed. No one measure can resolve the language test's reliability, specifically the speaking competence exam [32].

2. Literature Review

The literature review is carried out under two distant variables of the study—i.e., nature of testing speaking and reliability.

2.1. Oral Testing. Oral ability testing as an element of English instruction is a necessary procedure, not just because it provides a valuable platform for data on the effectiveness of education [33, 34]. Additionally, it could facilitate and expedite instruction, enhance learners' motivation to improve their language proficiency, and strengthen the evaluation process [35, 36]. The assessment of speaking ability has been viewed as a prominent issue in the language testing system, as speaking ability plays an essential part in language development and learning and has assumed a vital role in language education with the onset and emphasis on communicative language teaching in remote or online learning background. Speaking ability is embedded in culture, and "situation-based activity" is a significant component of daily life scenarios [37]. An ESL or EFL assessment is commonly considered more difficult than assessing other abilities, skills, or correctness ([38–40]).

Speaking tests cover various language learning areas, including vocab, proper grammatical usage, fluency, correctness, interaction, the social side of speaking, and task fulfillment [41, 42]. Additionally, assessing speaking is complicated because of its dynamic character, spontaneity, and appropriateness [11, 43–45]. To accomplish this, instructors, learners, and assessors must have a firm grasp of the features and structure of oral language that set it apart from other modes of language assessment [46, 47]. Ockey [48] asserts that Clark and Swinton established a theoretical framework for classifying three types of speech assessments: "direct, semidirect, and indirect exams." The direct and semidirect examinations need learners to present before assessors and discuss the assigned topic. At the same time, the indirect tests are part of the testing system's "pro-communicative" period and do not need learners to engage in communicative skills [49–51].

The oral assessment is among the most often utilized test types for evaluating speaking ability and substantially impacts language assessment. It is conducted with a single test taker and one or two qualified assessors or raters who assess or record their speaking ability on the predefined scale. It starts with introducing the individual, a warm-up chat to establish rapport, and then predetermined test tasks such as narrating an experience, an event, role-plays, or reversal interview. The majority of language assessments are semistructured. The IELTS speaking section is a critical

component of this speaking assessment, approved in over 100 countries worldwide [52]. The interview form of assessment enables the assessor or examiner to gain a holistic impression of the learners' speaking ability and compensate for the inadequacies of other elements of the language assessment process. Furthermore, it is pretty simple to train examiners and achieve good interrater reliability [53].

Another type of speaking assessment is the pair or group assessment. This evaluation method involves one or more assessors assessing the examinees' speaking performance in groups or couples. The paired test is used to evaluate large-scale speaking ability. Speaking evaluations emphasize the interaction between participants and test takers [54]. This enables a more flexible interaction among test takers and assessors and a broader type of discourse than formal interviews [55]. Both forms provide raters with handouts and speaking evaluation criteria. The speaking test is graded holistically or analytically, regardless of the type of communication.

2.2. Reliability. Reliability is an essential factor of every test. The goal of reliability is to assure the precision with which examinees' knowledge and performance are validated. The extent to which a test tool produces steady and consistent results is called its reliability [56]. The term "reliability" is defined as "the consistency of assessment" [57]. Thus, reliability argues that the findings are the most accurate and complete representation of a test participant's competency. This statement asserts that grading should be congruent with the test's or rater's reliability. The reliability of a test is characterized by its capacity to reflect the correctness and consistency of an evaluation. Traditionally, during the testing protocol, two reliability components are considered: interrater and intrarater reliability. Jeyaraman et al. [58] asserted that interrater reliability refers to the precision of grades provided by evaluators.

In contrast, intrarater reliability refers to the consistency of a rater's rating on distant times. This emphasizes that interrater consistency is established by comparing the grades assigned by various examiners. In contrast, intrareliability is found by evaluating the scores given by the same assessors for the same respondents over time. This demonstrates that there is no one-size-fits-all method for determining the reliability of an exam or test. Rater reliability is a concern because it incorporates individual subjectivity, affecting the marks assigned to various learners [59].

When it comes to assessing language learning's productive skills, the function of raters is always critical when it comes to determining practical ability. The reliability of an oral examination is rugged and necessitates remote measurements. Due to the subjectivity of the speaking assessment, some raters may be more moderate than others, affecting the reliability [60]. This is due to the rater's cultural context or present mood. The familiarity with the accent of the examinee may also influence the rater to award higher grades in the pronunciation part of the test [61].

Similarly, when a rater is familiar with L1 communication, they are more tolerant of granting respondents better scores. This demonstrates that the speaking exam scores are influenced in various ways [62–64]. Additionally, the degree to which raters' judgments contradict one another depends on the assessment scale, rubrics, and marking standards employed in a particular oral test. Because of the comprehension of the grading system, this rating criterion could have an impact on the intrarater reliability of the results. As a result, raters' knowledge of the grading scale and awareness of the rubric are also important in determining reliability.

Several studies in language assessment have investigated numerous components of speaking evaluation. Kang et al. [21] state that the research outcomes contribute significantly to understanding the speaking assessment concept. Several researchers have examined the speaking test's reliability. Nicholson [65] indicated that the speaking assessment was exceedingly consistent, but the validity argument appeared to be erroneous. The Khan et al. [66] study discovered discrepancies in examiners' scores. Further investigation revealed that the differences in the evaluators' scores were primarily due to one of the evaluators awarding scores in the grammar and vocabulary section of the test. Further, it could be enhanced by training raters before the implementation of the test.

Iwashita and Vasquez [67] and Benyo and Kumar [68] also investigated a speaking competency test format to develop a scale for ESP grading. The study found that specific aspects of the test, including fluency and vocabulary, had a persistent effect on the total scores provided by the assessors. The findings of this study are expected to have a potential influence on the construction of scales. Likewise, Demirel and Baser [50] discovered that assessing the reliability of speaking skills is not an easy process because it is influenced by various factors, including the test's construct, task, and understanding of the learners' background. Numerous studies have been conducted to verify the IELTS speaking test's reliability [24, 42, 52, 69]. According to research, most of the IELTS speaking tests are accurate and consistent. The IELTS speaking test is considered valid regarding the content covered, accessibility, and presentation.

On the other hand, the researchers concentrated on the introduction of two reviewers in the IELTS speaking test. The review of literature suggests that reliability analysis of the speaking test is negated by the scholars, and online assessment has recently emerged. Therefore, the purpose of this study is to measure the reliability of an oral examination that involves two raters evaluating a test taker concurrently by using the blackboard platform. The present study tries to answer the following research question: How reliable is a speaking test used for online assessment of Saudi EFL learners?

3. Methods

This attempt is aimed at evaluating the reliability of oral performance tests. Quantitative research design for data collection was utilized to answer the research question. The data for the measure of reliability were gathered using a speaking test devised by the exam committee. The test consists of six to eight tasks, each with a distinct set of questions.

Without knowledge of the tasks' contents, participants could choose themes for speaking at random. After selecting the task, learners have shared the task through the BB platform. They were allotted two minutes to read and think about the given task and were permitted to choose another task if they wanted to. Following some warm-up questions, participants were expected to describe the individual tasks, and the procedure was interactive.

The study included 56 CFY undergraduates who spent their first two semesters studying English language skills as a condition for admission to their majors. The participants were aged from 16 to 20 years. Each participant was a male student. The test was administered before the final exam of the first term. According to their entrance test, all participants had the same English proficiency. The participants were explained about the test procedure and given online training for the test, as for the first regular students had an online speaking exam. They received a BB-link to join a group; then, one of the evaluators had to split them into groups on BB. Finally, from the group, students were randomly invited to the main room of the BB for the speaking test.

Eight raters who are regular faculty members of the CFY faculty were engaged in the speaking test's scoring method. Since 2014, all evaluators have been conducting this type of test. Additionally, they underwent training courses for the speaking assessment. They are part of the regular staff of the CFY program and hold a master's degree in English and a CELTA teaching credential. The evaluators were between 34 and 56 years. The rating technique was conducted in pairs, with one participant and two raters in an online platform, and participants were assigned marks on the holistic approach.

The data collection instrument was a speaking test and the student's grades. The exam committee developed this test following the Cambridge University A-2 level speaking assessment criteria (English, 2011). The test consists of a variety of tasks selected from the course content. Each task takes between 8 and 16 minutes to complete. The overall score for the assessment was 15 points, with five points assigned to each of the three dimensions of the speaking test: task fulfillment, fluency and accuracy, and vocabulary use. Evaluators were supplied with each student's speaking criteria, rubrics, and rating form.

4. Data Analysis

Generally, Kuder Richardson and statistical correlation measurements are used to evaluate the test's reliability. Test/retest split-half technique and parallel form are used to determine the test's reliability. Syahidah and Umasugi [70] assert that conventional methods of reliability calculation have little relevance to oral examination since they are developed for a fixed number of preplanned topics and questions. Practical estimation for the speaking test assessment can be obtained by comparing raters' results to those of other raters with special measures. The interrater reliability was used to evaluate the speaking test's reliability for the current test. The overall interrater reliability was 0.70 for the speaking test. According to Hiser et al. [71], rater reliability can be assessed using correlation, regression, and the Bland-Altman test. To this aim, two measures were utilized to determine reliability: Bland-Altman and correlation. SPSS 22 was applied to conduct the analyses for both tests.

The findings are reported in the following stages to estimate the spoken proficiency test's reliability. Participants were based on evaluating 15 speaking test scores from both raters who assessed them concurrently. The first stage evaluated the test's interrater reliability. Due to the human component of the test procedure, two different tests were used to determine the speaking test's reliability. Because this is an assessment of productive ability testing, the rater's decision to assign marks may impact the speaking testing process. Interrater reliability was determined by employing correlation coefficients derived in SPSS software on all evaluators' marks. The evaluators were divided into four pairs to calculate the correlation, t. Interrater reliability is summarized in Table 1.

Table 1 presents the interrater reliability of the eight raters in four pairs. For each of the four couples, a correlation coefficient was generated. Eight raters were paired in four pairs for the data analysis. Correlation coefficients of the evaluators' ratings were 0.710, 0.600, 0.610, and 0.640 for four pairs. Correlation coefficients for the 1st pair were 0.710; the 2nd pair was 0.660, the 3rd pair was 0.610, and the 4th pair was 0.640. The first pair has an adequate level of reliability. However, the 2nd, 3rd, and 4th pairs have a fairly low level of reliability. Despite the 2nd and 3rd pairs' low reliability, the p values for all pairs were $p = 0.01$ which is significant.

4.1. Bland-Altman Test. The Bland-Altman test is done to evaluate the degree of agreement among raters. The raters' ratings were combined in 3 groups to determine interrater reliability. Figure 1 depicts pair 1 and 2 agreement.

Figure 1 depicts the consistency between raters 1 and 2. As illustrated in Figure 1, most points are located between the average value and zero, indicating that the raters are in agreement. When more than 50% of the points are close to zero, this implies that the raters are in agreement. Additionally, the average value of pair 1 and pair 2 is close to +1.96 SD and -1.96 SD, respectively. SD values for pair one and pair 2 are 1.26 and -1.03, respectively, which are within the acceptable norm of data to demonstrate agreement. The agreement between the raters' scores for pair three and pair four is depicted in Figure 2.

Figure 2 illustrates the agreement between raters C and D. Further, the chart demonstrates that most dots are located near the average value and zero lines, indicating that the raters agree. Likewise, the mean values of pair 3 and pair 4 are close to +1.96 SD and -1.96 SD, respectively. SD values for pair three and pair 4 are 1.60 and -1.31, respectively, which are within the usual norm of data and demonstrate agreement. The rater agreement of pairs 5 and 6 is depicted in Figure 3.

Figure 3 illustrates the agreement between raters 5 and 6. As shown from Figure 3, most of the dots are close to the average value and zero lines, indicating that the evaluators agree. If more than 50% of the scores are close to zero, this

Table 1: Correlations in pairs.

Pair 1		Rater 1	Rater 2
Rater 1	Pearson correlation	1	0.701 (∗∗)
	Sig. (2-tailed)	56	0.00
	N	0.710 (∗∗)	56
Rater 2	Pearson correlation	0.00	
	Sig. (2-tailed)	56	
	N		
Pair 2		Rater 1	Rater 2
Rater 1	Pearson correlation	1	0.660 (∗∗)
	Sig. (2-tailed)	56	0.00
	N	0.660 (∗∗)	56
Rater 2	Pearson correlation	0.00	
	Sig. (2-tailed)	56	
	N		
Pair 3		Rater 1	Rater 2
Rater 1	Pearson correlation	1	0.610 (∗∗)
	Sig. (2-tailed)		0.00
	N	56	56
Rater 2	Pearson correlation	0.6100 (∗∗)	
	Sig. (2-tailed)	0.00	
	N	56	
Pair 4		Rater 1	Rater 2
		1	0.640 (∗∗)
Rater 1	Pearson correlation		0.00
	Sig. (2-tailed)	56	56
	N		
Rater 2	Pearson correlation	0.6400 (∗∗)	
	Sig. (2-tailed)	0.00	
	N	56	

indicates that the raters are in agreement. Also, the mean values of pair 5 and pair 6 are close to +1.96 SD and -1.96 SD, respectively. SD estimation for pairs 5 and 6 are 1.67 and -1.31, respectively, which are substantially within the acceptable norm of data to demonstrate agreement.

5. Discussion and Recommendation

Our conclusions are based on the oral data. We used reliability guidelines, such as rubrics and two examiners, to minimize the impact of human involvement in the testing system. When evaluating language learners' productive abilities, the role of raters is always essential to determine practical possibility. Oral examinations are highly unreliable, requiring the use of virtual measurements. Because the speaking assessment is subjective, some evaluators may be more moderate than others, impacting the extent of reliability. The literature review indicates that speaking evaluation was examined from various perspectives, emphasizing broad subject areas: speaking capability structures, rater impacts, factors affecting spoken efficiency, test design, test score generalization, assessing scale assessment, and test utilized. The vital aspect which impacts evaluation has been overlooked. The present study sheds some light to add to the literature to offer some insight into reliability analysis. The study offers insight for language scholars by presenting a way to check the reliability of the speaking test.

The speaking test's reliability was tested in two methods. The correlation coefficient suggested that the rater's interrater reliability is insufficient to satisfy the intended standard of test reliability. Nevertheless, the first pair's reliability was 0.710, which is deemed satisfactory for the online speaking test, but the reliability of the 2nd, 3rd, and 4th pairs is assessed 0.610, 0.600, and 0.640, respectively, which is not satisfactory. The discrepancy in the reliability estimation may be the result of the online assessment where an evaluator cannot see the confidence and facial expression of the participants. Another attribution for the low reliability can be the informal way of testing. Although the pairs' reliabilities were insufficient, the p values for all four pairs were less than $p = 0.00$, less than 0.05. This demonstrates the reliability of the speaking exam utilized at CFY. The gap in the

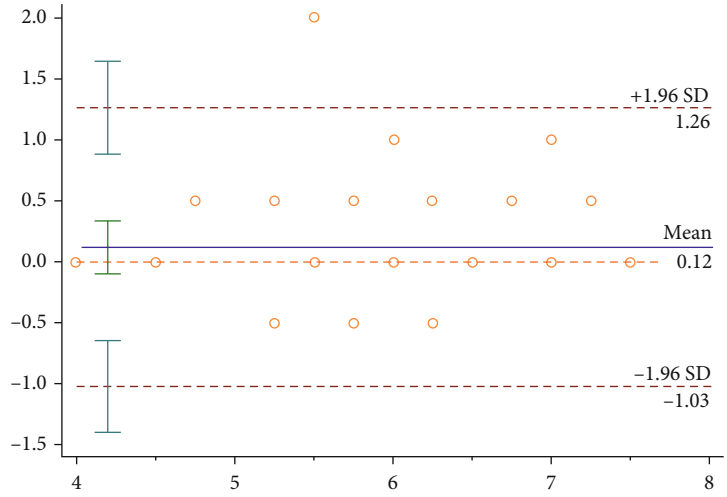

Figure 1: Mean of pairs 1 and 2.

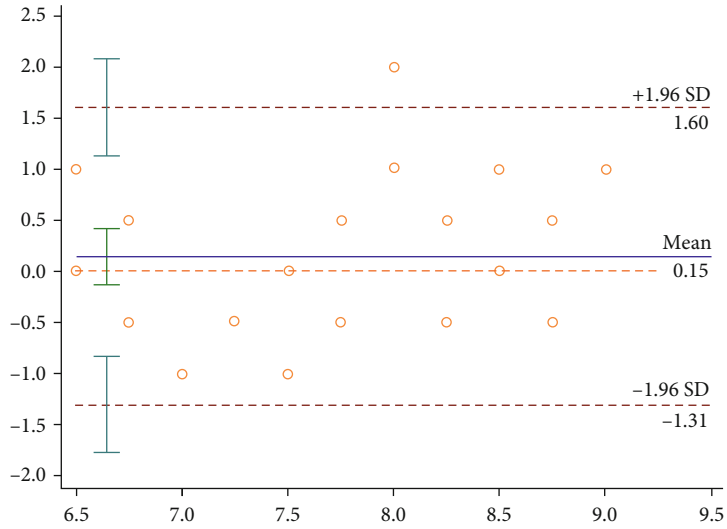

Figure 2: Mean of pairs 3 and 4.

interrater reliability findings could be because the correlation identifies how many identical scores were assigned to respondents, which is not achievable when scores are given in point and more significant than zero.

This concludes by using the Bland-Altman test, which determines the agreement between two raters. This could be a valuable way of gauging the reliability of the test, particularly in speaking skill research. Bland-Altman analysis revealed that all four pairs of evaluators have the interrater agreement. The data points are more equidistant from the zero lines. When more than 50% of the scores are close to zero, this indicates that the raters are in agreement. This was evident in each of the four pairs. Likewise, the Bland-Altman mean values were close to +1.96 and -1.96 in all three figures. Hence, it may be stated that the CFY speaking test is reliable and can present a good evaluation. Assessing the reliability of speaking ability is not an easy process, as it is influenced by various factors, including the test's structure, task, and knowledge of the participants' background. The findings assert that instead of using the correlation coefficient test to determine the reliability. The Bland-Altman test is more suitable for oral examination. The correlation test measures the degree of identical scoring, and hence, in speaking evaluation, there is no one or zero scoring; this leads to the use of the Bland-Altman test in virtual and face-to-face testing.

The research findings are partly consistent with those of [66], who suggest that these analysis results are instrumental in predicting the test's reliability. The present attempt also observed some consistency with O'Mahony [57] findings, who investigated the reliability of an oral test. The study's results revealed that the spoken test was highly reliable; yet, the reliability in this study seemed to meet the established standard of reliability. This could result from the various concerns, including remote assessment and raters assigned point values, resulting in a lesser level of reliability.

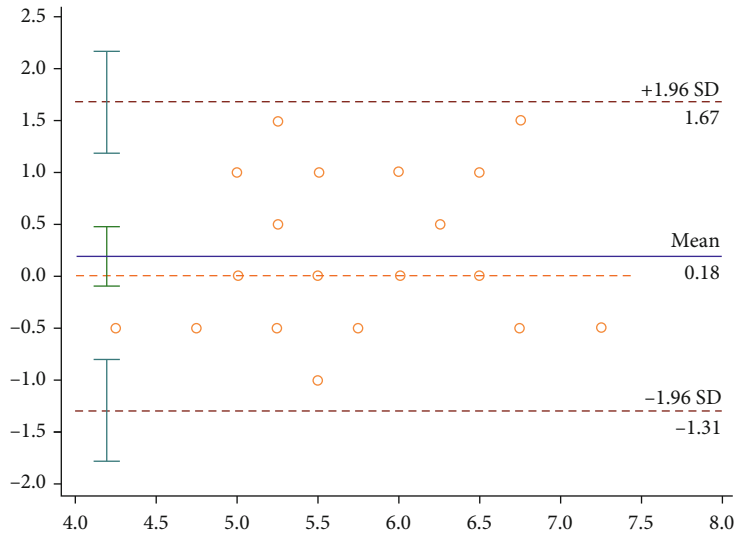

Figure 3: Mean of pairs 5 and 6.

The findings are congruent with Iwashita and Vasquez [67] investigation, which demonstrated inconsistencies in examiners' assessments. The differences in the raters' ratings were primarily caused by one of the raters awarding high marks for grammar and vocabulary usage. Furthermore, the findings are consistent with Iwashita and Vasquez [67], who analyzed various spoken proficiency tests to develop a rating scale for ESP. The results indicated that specific aspects of the test, including fluency and vocabulary, had a persistent effect on the total scores provided by the evaluators. Therefore, the findings corroborated the findings of previous studies [24, 42, 52, 69] that established the IELTS speaking test's reliability. According to statistics presented in the studies, most IELTS speaking tests are accurate and reliable.

Online learning and evaluation are always challenging. Learners need the motivation to participate in the learning and testing procedure. The study could be expanded in a variety of ways. The number of raters can be increased, and pairs of participants can be swapped for grading purposes. Also, the rater training provision before the test administration can result in a different outcome. The reliability of the raters for speaking skills revealed some detrimental differences among the rate end. It would also be beneficial if the grading system is made more transparent to the evaluator, contributing to the test's reliability. Finally, in online assessment, the rater may ask the participants a role rehearsal; this will help evaluators gain an accurate picture of the speaking proficiency. Moreover, learners should also be given training in an online way to understand how the speaking test is carried out in remote learning.

6. Implications and Limitations

The study examined and reviewed the reliability analysis of the speaking test. This study concludes that using the Bland-Altman test can help teachers and scholars determine the test's reliability. As oral examination includes human interaction, it is not feasible to agree on 1 or 0 points. To this end, researchers, examiners, and test developers can use the Bland-Altman test to check the reliability of the speaking test, which determines the degree of agreement between two raters. This could be a valuable way of gauging the reliability of the test, particularly in speaking skill research. The study findings can also help the research scholar in oral or spoken skill development.

Although the subject matter of speaking frameworks has garnered considerable research interest in the field, as illustrated by the interpretation finding of this research, it appears that there is still a long way from attaining a detailed and perfectly alright comprehension of determining the reliability of speaking ability. The study had some limitations. First, the study was limited to a only campus and one level of the students; future studies are operative to include participants for distant institutions to present more generalized findings. Moreover, the study includes only male participants and the sample was small too. The inclusion of both genders may present different findings.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] A. Al-Ansi, H. G. Olya, and H. Han, "Effect of general risk on trust, satisfaction, and recommendation intention for halal food," *International Journal of Hospitality Management*, vol. 83, pp. 210–219, 2019.

[2] H. Al-Nofaie, "Saudi University students' perceptions towards virtual education during COVID-19 pandemic: a case study of

language learning via blackboard," *Arab World English Journal*, vol. 11, no. 3, pp. 4–20, 2020.

[3] J. Thongbunma, P. Nuangchalerm, and S. Supakam, "Secondary teachers and students' perspectives towards online learning amid the COVID-19 outbreak," *Gagasan Pendidikan Indonesia*, vol. 2, no. 1, pp. 1–9, 2021.

[4] M. Z. Hoq, "E-learning during the period of pandemic (COVID-19) in the kingdom of Saudi Arabia: an empirical study," *American Journal of Educational Research*, vol. 8, no. 7, pp. 457–464, 2020.

[5] R. M. I. Khan, N. Radzuan, S. Farooqi, M. Shahbaz, and M. Khan, "Learners' perceptions on WhatsApp integration as a learning tool to develop EFL spoken vocabulary," *International Journal of Language Education*, vol. 5, no. 2, pp. 1–14, 2021.

[6] O. B. Adedoyin and E. Soykan, "COVID-19 pandemic and online learning: the challenges and opportunities," *Interactive Learning Environments*, vol. 1-13, pp. 1–13, 2020.

[7] A. Kukulska-Hulme, H. Lee, and L. Norris, *Mobile Learning Revolution*, The Handbook of Technology and Second Language Teaching and Learning, 2017.

[8] M. Ajmal and T. Kumar, "Using DIALANG in assessing foreign language proficiency: the interface between learning and assessment," *Asian ESP Journal.*, vol. 16, no. 2.2, pp. 335–362, 2020.

[9] M. S. Bacha, T. Kumar, B. S. Bibi, and M. M. Yunus, "Using English as a lingua franca in Pakistan: influences and implications in English language teaching (ELT)," *Asian ESP Journal*, vol. 17, no. 2, pp. 155–175, 2021.

[10] J. Jenkins, W. Baker, and M. Dewey, *The Routledge Handbook of English as a Lingua Franca*, Routledge, 2017.

[11] T. Kumar, "Social networking sites and grammar learning: the views of learners and practitioners," *International Journal of Early Childhood Special Education (INT-JECSE)*, vol. 13, no. 2, pp. 215–223, 2021.

[12] S. Zuparova, A. Shegay, and F. Orazova, "Approaches to learning English as the source of all," *European Journal of Research and Reflection in Educational Sciences*, vol. 8, no. 5, 2020.

[13] A. Nugroho and A. E. P. Atmojo, "Digital learning of English beyond classroom: Efl learners' perception and teaching activities," *JEELS (Journal of English Education and Linguistics Studies)*, vol. 7, no. 2, pp. 219–243, 2020.

[14] R. M. I. Khan, G. Mustafa, and A. A. Awan, "Learners' attitudes on the infusion of cooperative learning in education," *Orient Research Journal of Social Sciences*, vol. 5, no. 2, 2018.

[15] M. Shahbaz and R. M. I. Khan, "Use of mobile immersion in foreign language teaching to enhance target language vocabulary learning," *MIER Journal of Educational Studies Trends & Practices*, pp. 66–82, 2017.

[16] S. S. Alotaibi and T. Kumar, *Promoting teaching and learning performance in mathematics classroom through e-learning*, Opción, Año 35, Especial No.19, 2019.

[17] F. Çakmak, E. Namaziandost, and T. Kumar, "CALL-enhanced L2 vocabulary learning: using spaced exposure through CALL to enhance L2 vocabulary retention," *Education Research International*, vol. 2021, Article ID 5848525, 8 pages, 2021.

[18] S. Sreena and M. Ilankumaran, "Developing productive skills through receptive skills – a cognitive approach," *Engineering and Technology*, vol. 7, no. 4.36, pp. 669–673, 2018.

[19] R. M. I. Khan, N. R. M. Radzuan, M. Shahbaz, A. H. Ibrahim, and G. Mustafa, "The role of vocabulary knowledge in speaking development of Saudi EFL learners," *Arab World English Journal (AWEJ)*, p. 9, 2018.

[20] S. Witzigmann and S. Sachse, *Diagnostic Competencies of Prospective Teachers of French as a Foreign Language: Judgement of Oral Language Samples: RISTAL*, 2021.

[21] O. Kang, D. Rubin, and A. Kermad, "The effect of training and rater differences on oral proficiency assessment," *Language Testing*, vol. 36, no. 4, pp. 481–504, 2019.

[22] O. Kang and A. Kermad, *Assessment in Second Language Pronunciation*, Routledge, 2017.

[23] M. Polat, "A rasch analysis of rater behaviour in speaking assessment," *International Online Journal of Education and Teaching (IOJET)*, vol. 7, no. 3, pp. 1126–1141, 2020.

[24] P. Seedhouse and F. Nakatsuhara, *The Discourse of the IELTS Speaking Test: Interactional Design and Practice*, Cambridge University Press, 2018.

[25] S. J. Youn, "Rater variability across examinees and rating criteria in paired speaking assessment," *Papers in Language Testing and Assessment*, vol. 7, no. 1, pp. 32–60, 2018.

[26] G. J. Ockey, D. Koyama, E. Setoguchi, and A. Sun, "The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students," *Language Testing*, vol. 32, no. 1, pp. 39–62, 2015.

[27] H. Cai, *Distinguishing Language Ability from the Context in an EFL Speaking Test another Generation of Fundamental Considerations in Language Assessment*, Springer, 2020.

[28] R. Hughes and B. S. Reed, *Teaching and Researching Speaking*, Routledge, 2016.

[29] J. Khan, K. H. Yuen, B. H. Ng et al., "Bioequivalence evaluation of two different controlled release matrix formulations of ketoprofen tablets in healthy Malaysian volunteers," *Latin American Journal of Pharmacy*, vol. 30, no. 10, p. 1991, 2011.

[30] S. G. Sireci and M. Faulkner-Bond, "Promoting validity in the assessment of English learners," *Review of Research in Education*, vol. 39, no. 1, pp. 215–252, 2015.

[31] N. Sultana, "Language assessment literacy: an uncharted area for the English language teachers in Bangladesh," *Language Testing in Asia*, vol. 9, no. 1, pp. 1–14, 2019.

[32] R. Nair, G. Tsakos, and R. Yee Ting Fai, "Testing reliability and validity of oral impacts on daily performances for Chinese-speaking elderly Singaporeans," *Gerodontology*, vol. 33, no. 4, pp. 499–505, 2016.

[33] R. Esmaili, S. M. Mousavi-Davoudi, and F. Nasiri-Amiri, "The impact of spiritual intelligence on aggressive behavior, considering the mediating role of professional ethics: a case study of nurses of Imam Ali (pbuh) hospital in Alborz, Iran," *Journal of Pizhūhish dardīn va salāmat*, vol. 7, no. 3, pp. 35–50, 2021.

[34] E. Namaziandost, L. Neisi, Kheryadi, and M. Nasri, "Enhancing oral proficiency through cooperative learning among intermediate EFL learners: English learning motivation in focus," *Cogent Education*, vol. 6, no. 1, p. 1683933, 2019.

[35] L. Davis, V. Timpe-Laughlin, L. Gu, and G. Ockey, *Interactive Speaking Tasks Online*, Useful Assessment and Evaluation in Language Education, 2018.

[36] R. Saad, G. Murugiah, J. Abdulhamid, E. Yusuf, and M. Fadli, "Comparative study between percolation and ultrasonication for the extraction of hibiscus and jasmine flowers utilizing antibacterial bioassay," *International Journal of*

Pharmacognosy and Phytochemical Research, vol. 6, no. 3, pp. 472–476, 2014.

[37] H. Lee, S. Lee, J. Ko, and H. Bang, "Investigating the effects of course satisfaction and career decision-making efficacy on intrinsic motivation of undergraduates in beauty health major," *Educational Sciences: Theory and Practice*, vol. 21, no. 3, pp. 147–157, 2021.

[38] S. Bal-Taştan, S. M. M. Davoudi, A. R. Masalimova et al., "The impacts of teacher's efficacy and motivation on student's academic achievement in science education among secondary and high school students," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 14, no. 6, pp. 2353–2366, 2018.

[39] M. Mohammadi and N. Danesh Pouya, "The relationship between EFL learners' mental toughness and critical thinking," *Journal of Social science and Humanities Research*, vol. 9, no. 2, pp. 25–36, 2021.

[40] J. Ranalli, S. Link, and E. Chukharev-Hudilainen, "Automated writing evaluation for formative assessment of second language writing: investigating the accuracy and usefulness of feedback as part of argument-based validation," *Educational Psychology*, vol. 37, no. 1, pp. 8–25, 2017.

[41] K. Fartash, S. M. M. Davoudi, T. A. Baklashova et al., "The impact of technology acquisition & exploitation on organizational innovation and organizational performance in knowledge-intensive organizations," *Eurasia Journal of Mathematics, Science and Technology Education*, vol. 14, no. 4, pp. 1497–1507, 2018.

[42] J. Li, "An evaluation of IELTS speaking test," *Open Access Library Journal*, vol. 6, no. 12, pp. 1–17, 2019.

[43] R. Ahmed and A. Al-Kadi, "Online and face-to-face peer review in academic writing: frequency & preferences," *Eurasian Journal of Applied Linguistics*, vol. 7, no. 1, pp. 169–201, 2021.

[44] Ö. Ardıç and H. Çiftçi, "ICT competence and needs of Turkish EFL instructors: the role of gender, institution and experience," *Eurasian Journal of Applied Linguistics*, vol. 5, no. 1, pp. 153–173, 2019.

[45] J. Fan and X. Yan, "Assessing speaking proficiency: a narrative review of speaking assessment research within the argument-based validation framework," *Frontiers in Psychology*, vol. 11, p. 330, 2020.

[46] Z. Qinghua, "Evaluation and prediction of sports health literacy of college students based on artificial neural network," *Revista De Psicología Del Deporte (Journal of Sport Psychology)*, vol. 30, no. 3, pp. 9–18, 2021.

[47] A. B. P. Sari, "WhatsApp-based speaking test in EFL context," *Journal of English Language Studies*, vol. 5, no. 2, pp. 175–188, 2020.

[48] G. J. Ockey, "Oral language proficiency tests," *The TESOL Encyclopedia of English language teaching*, vol. 1-5, 2018.

[49] S. A. Ahmadi, M. Z. Sakha, S. Ebadi, and A. K. Panda, "Study of milk and dairy products Staphylococcus contamination and antimicrobial susceptibility sold in local markets around Kabul University," *International Journal of Innovative Research and Scientific Studies*, vol. 4, no. 1, pp. 20–24, 2021.

[50] E. T. Demirel and Z. Baser, *Examination of Speaking Test Performance in Structured Group Tasks: An Interactional Perspective Design Solutions for Adaptive Hypermedia Listening Software*, IGI Global, 2021.

[51] M. A. Rasuli and S. Torii, "Feasibility of solar air conditioning system for Afghanistan's climate," *International Journal of Innovative Research and Scientific Studies*, vol. 4, no. 2, pp. 120–125, 2021.

[52] C. J. Fernandez, "Behind a spoken performance: test takers' strategic reactions in a simulated part 3 of the IELTS speaking test," *Language Testing in Asia*, vol. 8, no. 1, pp. 1–20, 2018.

[53] L.-F. Huang, S. Kubelec, N. Keng, and L.-H. Hsu, "Evaluating CEFR rater performance through the analysis of spoken learner corpora," *Language Testing in Asia*, vol. 8, no. 1, pp. 1–17, 2018.

[54] K. Frost, J. Clothier, A. Huisman, and G. Wigglesworth, "Responding to a TOEFL iBT integrated speaking task: mapping task demands and test takers' use of stimulus content," *Language Testing*, vol. 37, no. 1, pp. 133–155, 2020.

[55] C. Roever and G. Kasper, "Speaking in turns and sequences: interactional competence as a target construct in testing speaking," *Language Testing*, vol. 35, no. 3, pp. 331–355, 2018.

[56] L. Cohen, L. Manion, K. Morrison, L. Cohen, L. Manion, and K. Morrison, *Validity and Reliability Research Methods in Education*, Routledge, 2017.

[57] C. O'Mahony, "Reliability and validity in language testing–a real conflict?," *Center for English Language Education (CELE)*, vol. 27, pp. 133–140, 2019.

[58] M. M. Jeyaraman, N. Al-Yousif, R. C. Robson et al., "Inter-rater reliability and validity of risk of bias instrument for non-randomized studies of exposures: a study protocol," *Systematic Reviews*, vol. 9, no. 1, pp. 1–12, 2020.

[59] R. Schoonen, *Measurement Generalizability: Considerations on Reliability and Validity the Routledge Handbook of Second Language Acquisition and Language Testing*, Routledge, 2020.

[60] J. Blais, A. E. Forth, and R. D. Hare, "Examining the interrater reliability of the Hare Psychopathy Checklist—revised across a large sample of trained raters," *Psychological Assessment*, vol. 29, no. 6, pp. 762–775, 2017.

[61] M. S. Park, "Rater effects on L2 oral assessment: focusing on accent familiarity of L2 teachers," *Language Assessment Quarterly*, vol. 17, no. 3, pp. 231–243, 2020.

[62] V. Bogorevich, *Native and Non-Native Raters of L2 Speaking Performance: Accent Familiarity and Cognitive Processes*, Northern Arizona University, 2018.

[63] H. Lee, "The effects of raters familiarity with test takers L1 in assessing accentedness and comprehensibility of independent speaking tasks," *SNU Working Papers in English Linguistics and Language*, vol. 15, 2017.

[64] K. Zhao, *Investigating the Effects of Rater's Second Language Learning Background and Familiarity with Test-Taker's First Language on Speaking Test Scores*, Brigham Young University, 2017.

[65] S. J. Nicholson, "Evaluating the TOEIC® in South Korea: practicality, reliability and validity," *International Journal of Education*, vol. 7, no. 1, pp. 221–233, 2015.

[66] R. M. I. Khan, N. R. M. Radzuan, T. Kumar, and M. Shahbaz, "An investigation of the reliability analysis of speaking," *The Asian EFL Journal*, vol. 27, no. 3, pp. 356–373, 2020.

[67] N. Iwashita and C. Vasquez, "An examination of discourse competence at different proficiency levels in IELTS speaking part 2," *IELTS Research Reports Online Series*, vol. 5, pp. 1–44, 2015.

[68] A. Benyo and T. Kumar, "An analysis of Indian EFL learners' listening comprehension errors," *Asian ESP Journal*, vol. 16, no. 5.2, pp. 69–85, 2020.

[69] F. Nakatsuhara, C. Inoue, V. Berry, and E. D. Galaczi, *Exploring Performance across Two Delivery Modes for the IELTS Speaking Test: Face-to-Face and Video-Conferencing Delivery (Phase 2)*, IELTS Partners, 2017.

[70] U. Syahidah and F. Umasugi, "A design of speaking assessment rubric for English immersion camp," *Exposure: Jurnal Pendidikan Bahasa Inggris*, vol. 10, no. 1, pp. 31–46, 2021.

[71] S. Hiser, C. R. Chung, A. Toonstra et al., "Inter-rater reliability of the Johns Hopkins Highest Level of Mobility Scale (JH-HLM) in the intensive care unit," *Brazilian Journal of Physical Therapy*, vol. 25, no. 3, pp. 352–355, 2021.