An Investigation of the Reliability Analysis of Speaking Test

Raja Muhammad Ishtiaq Khan[1], Noor Raha Mohd Radzuan[2], Muhammad Shahbaz[3] & Tribhuwan Kumar [4]

[1] English Language Lecturer, Majma'ah University, Zulfi, Kingdom of Saudi Arabia
https://orcid.org/0000-0003-0873-4124, E-mail: r.khan@mu.edu.sa

[2] Center for Modern Languages, Universiti Malaysia Pahang, Kuantan, Pahang, Malaysia
http://orcid.org/0000-0001-7418-6360, E-mail: nraha@ump.edu.my

[3] Department of English, GC Women University Sialkot, Sialkot, Pakistan,
E-mail: m.shahbaz@gcwus.edu.pk

[4]Assistant Professor of English, College of Science and Humanities at Sulail, Prince Sattam Bin Abdulaziz University, KSA, Email: t.kumar@psau.edu.sa

Bio-profiles:

Raja Muhammad Ishtiaq Khan is an English language Lecturer at Al-Majma'ah University, Saudi Arabia. He has a Cambridge CELTA certificate in teaching and has 10 years of experience in ESL and EFL. He is a Ph.D. scholar and his research interests include Applied Linguistics, MALL, L2 Vocabulary learning and Teaching, Second Language Acquisition, and EFL teaching.

Noor Raha Mohd Radzuan, Ph. D., is an Associate Professor and the Head of English Language Department at the Centre for Modern Languages and Human Sciences, University Malaysia Pahang. Her main interest is in Applied Linguistics research, specifically in second language oral communication, English for Specific Purposes, and English language teaching.

Muhammad Shahbaz is currently serving as an Assistant Professor at the Department of English, GC Women University Sialkot. He holds a Ph.D. Applied Linguistics and more than 10 years of working experience as an EFL educator. His research interests include SLA, L2 Motivation Research, Individual Differences in SLA, L2 Vocabulary development, Language Testing, and Evaluation.

Tribhuwan Kumar, Ph. D. is an Assistant Professor in the College of Science and Humanities, Sulail at Prince Sattam bin Abdulaziz University, Saudi Arabia, where he has been a faculty member since 2015. Before joining this university, he has taught in many institutions in India since 2010 including SRM University, NCR Campus, Ghaziabad. His research areas are British Literature, Indian English Literature, Applied Linguistics, discourse analysis, and other interdisciplinary subjects in language and literature.

Abstract

Speaking skill is marked as one of the most significant for language learning. The assessment of the oral test remained difficult because of the involvement of humans. The speaking assessment is based on the reliability of the test and reliability relies on the raters' score. To this, the present study is an attempt to estimate the inter-rater reliability of the Speaking test used in the Common First Year program (CFY) of a public university in Saudi Arabia. The data were collected through a scoring sheet from 62 EFL learners. 6 raters were involved in the rating of 61 participants of the study. Quantitative data analysis was used to estimate the reliability of the speaking test. The correlation coefficient and Bland-Altman test were used to measure the agreement between the raters. SPSS was used for data analysis in the present study. The result of the study indicated that the speaking test used in the CFY program is in the accepted norm of the reliability values.

Keywords: *Speaking, Reliability, EFL, Oral Test, raters*

Background

With the advancement of communicative instruction, the basic goal of speaking proficiency has gained more importance in language instruction and the ability to use the language appropriately in a social context has become the goal (BAUTISTA, Samonte, Improgo, Gutierrez, 2020; Alrefaee, Mudkanna & Almansoob, 2020). The value of communicative competence transcends speaking fluency and accuracy (Shariq, 2020; p. 236). As an outcome, testing speaking performance, particularly speaking capability has emerged as one of the key issues in the testing mechanism of language development. There are various restrictions because of the nature of the speaking skill. The fundamental issue in oral testing skills is the requirement to outline the tasks that structure an illustration of the sample of the population of the speaking tasks, explaining the outcome of that present the test takers' speaking capability. Likewise, many elements influence our notion of how better someone can express himself orally. As the nature of the oral skill itself is still infancy to define it precisely, there is an incongruity that allows assessing the various elements of the speaking skill.

In general, the speaking assessment is done based on pronunciation, vocabulary use, and grammatical accuracy. Similarly, the relevance and fluency are also common constituents of the speaking test. Due to the various components of the speaking test, its true assessment is not merely simple as compared to the other skills. Kemiläinen (2018) argue that in the process of the evaluation of oral skills there may be certain discrepancies as the test-taker has to use language in any way because of its collaborative nature. Besides, speaking evaluation mostly requires human evaluators. Hence the allotting scores of the speaking test are largely subjective biased. Silvia (2011) emphasizes this issue as the main problem in speaking evaluation, as the subjectiveness of the scoring process can direct the rater discrepancies or shifts creating an impact on test-takers' marks which can influence the rater reliability conversely. Thus, the scoring criterion is an essential element of the speaking test (Ling *et al.*, 2014; Tuan, 2012). Speaking evaluation has some practical limits too, which leads to inconsistent results. This involves the time, large numbers of the test-takers at the same time, administrative costs, rater's mood, training of the raters, testing length, use of rubrics, and the total time of the test. Despite these limitations, nowadays numerous schools, colleges, universities, and language

testing agencies are evaluating the test-takers' speaking proficiency. The speaking performance is being measured by using various types of tasks, including presentation, individual or group interviews, and role-plays which are anticipated to bring in the evidence on test-takers' speaking ability.

According to Vera & Vera (2018), oral communication has two skills or components. They are listening and speaking skills and performance in the speaking skills test is a major component-indicator of oral communication skills, the majority of them registered within the bracket of 'good' performance level. The evaluation process greatly depends on many aspects which can impede the learners' speaking performance. There is a need for well organized, researched, and documented account of the reliability and validity of the exam scores with observed logical evidence (Franklin *et al.*, 2001). The core point of the language testing should be focused on the precision of the analyses of learners' responses which can be justified on the constructs of the measures (Kernot *et al.*, 2015; Latief, 2016). The idea of language assessment is not simply on constructing the instruments for assigning certain levels or grades of the language proficiency but its to facilitate an outline of the categories of the testimony to be offered in case the accuracy of the interpretations of the proficiencies of the test marks are to be rationalized (Fulcher, 2013). Thus a speaking performance test should be supported by attaining the evidence to assist the aim that test is fulfilling the function that it is meant to be. This essentially includes offering data about distant validity measures together with authentic reliability. Nevertheless, the research indicates that there is a merely small part of the validity concern is met and there is no single measure that can fulfill the concern of the reliability of the language test particularly the speaking proficiency (Li, 2019).

Testing speaking

Testing oral proficiency as a component of teaching English, is an extremely significant process, not only as it can be a useful foundation of the data about the efficiency of the teaching and learning (Rohan-Minjares *et al.*, 2019). Moreover, it can be used to foster and expedite the teaching, stimulate the learners' motivation to develop their language proficiency as well as the development of the evaluation process (Ockey, 2018). The evaluation of the speaking performance has emerged as one of the key issues

in the testing system of the language as the speaking skill has the key role in the language learning and development and has attained the central position in the teaching and learning language with the emergence and focus on communicative language teaching. Speaking skill is a part of the society and “situation-based activity’ is a fundamental aspect of the daily lives situations (Namaziandost & Nasri, 2019). The evaluation of the foreign language or second language is often deemed to be a bit more challenging than the evaluation of the other skill, capabilities, or accuracy (Sari & Nike, 2017).

Testing speaking involves various aspects of language learning including vocabulary, the correct use of grammar, fluency, accuracy, interaction, the social aspect of speaking, and completion of the task (Bahrani & Soltani, 2011). Moreover, the evaluation of speaking is also difficult due to its dynamic nature, unpredictability, and comprehensibility (Bygate, 2009). To this end, the teachers, learners, and evaluators need to have a vivid understanding of the characteristics and nature of the oral language which distinguishes it for other forms of the language evaluation (Bygate, 2009; McCarthy & O'Keeffe, 2004).

Clark and Swinton (1979) mark a theoretical foundation to categorize three kinds of speaking evaluation, “direct, semi-direct and indirect tests". The direct test and semi-direct tests require learners to appear in front of the examiners and they have to speak on the allocated topic, whereas the indirect tests are the part of the “procommunicative" time in the testing system where the learners are not needed to take part in a speaking activity. Oral proficiency interview (OPI) is one of the most commonly used test formats of testing speaking skill and it has exerted a convincing influence on language testing. It is administered with a one test-taker and one or two trained evaluators or rater to evaluate or record the speaking performance on the given scale. It primarily begins with the introduction of the candidate, warming up a discussion to maintain the interaction followed by preformed test tasks including describing an event, picture or illustration, role play, or reverse interview. Most of the language interviews are semi-structured interviews. Speaking part of the IELTS test is one of the key forms of this type of speaking evaluation, which is accepted over 100 countries around the globe. Interview mode of assessing helps the rater or evaluator to get the overall sense of the speaking competence of the learners and can surpass the shortcoming of the other aspects of the language

evaluation process. Moreover, it is comparatively easy to train the examiners and gain high inter-rater reliability (Fulcher & Reiter, 2003).

Another form of the speaking test is testing in pairs or groups. In this form of the assessment, one or more evaluators access the test takers speaking proficiency either in pairs or small groups. The paired test is administered in testing large scale speaking proficiency. Interaction between the participants and test-takers is the major focus of both forms of speaking assessments. This offers a flexible mode of interaction between the test-takers and evaluators which further obtains a wider form of the discourse as compared to the conventional interview process (Dimitrova-Galaczi, 1969; May 2009, 2011). The raters in both formats are given the handouts of the speaking marking criteria. Speaking test is rated on the holistic or analytical rating scales depending on the nature of speaking proficiency.

Reliability

Reliability is regarded as one of the most important elements of any test. The purpose of the reliability is to ensure the accuracy of the ratification of the test-takers' knowledge and proficiency. Reliability measures the degree to which a test tool yields consistent and stable results (Golafshani, 2003). The conception of reliability is illustrated as "the consistency of measurement" (Bachman and Palmer, 1996). Reliability, thus, is an assertion that the results of the tests are the true and best possible indication of a test taker's proficiency. This asserts that the marking should be consistent with the test reliability of rater reliability. The core of the reliability of a test is that it indicates the accuracy and consistency of an assessment. Generally, two aspects of the reliability are contemplated during the testing procedure; inter-rater and intra-rater. According to Bachman *et al.* (2002), inter-rater reliability is related to the consistency of scores given by a group of raters and inter-rater reliability denotes how constant is the rating of a rater on different times. This asserts that inter-rater reliability is attained by comparison of scores awarded by the different examiners, whereas the intra-reliability is formed by comparing the scores of the same examiners for the same tests-takers at different periods. This indicates that there is no perfect and easy way that determines the reliability of a test.

Rater reliability is also challenging as it involves human subjectivity nature which affects the scores to different learners (Gwet, 2014).

In testing productive skills of language learning, the role of raters is always significant in testing writing and speaking skills. The reliability of an oral assessment demanding and requires distant measures. The subjective nature of the speaking evaluation can lead some raters to be lenient and some too strict which affects the reliability (Steenson *et al.*, 2013). This is because of the cultural background or the rater's mood. The intimacy of the test-takers accent made rater in giving higher scores for the pronunciation section (Carey *et al.*, 2011). Likewise, (Winke *et al.*, 2013) have argued that if a rater has the exposure of L1 communication, they tend to be lenient in awarding higher scores to the participants. This shows that the scores of the speaking test are affected in many ways. Moreover, the contradiction in raters' judgment also greatly based on the rating scales, rubrics used, and grading criteria. These rating criteria can also affect the intra-rater reliability because of the understanding of the grading system. So, the rater's knowledge of the grading system and rubric understanding is also crucial in governing reliability.

Numerous investigations in the testing language have already been carried out with the purpose to explore the various aspects of speaking assessment. Fujinaga *et al.* (2007) assert that the results of such analysis offer an important role in explaining the construction of the speaking evaluation. Some researchers have carried out the reliability of the speaking test. To begin with, Ozer *et al.* (2014) conducted the reliability of an oral test. The results of the study asserted that the speaking test was highly reliable, however, the construct of the validity appeared to invalid. Restrepo and Villa (2003) study indicated inconsistencies in the scores awarded by the examiners. The further analysis determined that the inconsistencies in the raters' scores have mainly resulted as one of the raters has award hiker's scores in grammar and vocabulary use. This can be improved by giving training to the raters.

Iwashita *et al.* (2008) also explored the form of speaking proficiency tests in order to establish a rating scale for ESP. The findings implied that certain features of the test had a resilient influence on the total scores assigned by the evaluators which include fluency and vocabulary. The outcomes of the study presented are contemplated to have a

persuasive association in the development of scales. Similarly, Li (2011) found that the assessment of the reliability of speaking proficiency is not an easy task and it is affected by many aspects including the construct of the test, the task of the test, the knowledge of the learners' background. Several investigations were made to determine the reliability of the IELTS speaking test (Karim & Haq, 2014; Li, 2019; Quaid, 2018; Read & Nation, 2006). The studies indicated that most of the IELTS speaking test is primarily valid and reliable. IELTS speaking test is regarded as reliable in terms of the contents used in the testing, accessibility, and appearance. However, the researchers focused on the inclusion of the two raters in the IELTS speaking test. The present study is an attempt to determine the reliability of an oral test which includes the two raters testing a test-taker at the same time. The present study aims to answer the following research question.

1. How reliable is the speaking test used for the common first-year students?

Methodology

The aim of the study is to determine the reliability of the speaking test. Therefore, the quantitative data collection method is employed to attend and analyse the data. A speaking test, developed by the administration of the CFY program was used to collect the data for reliability measure. The test contains 5 to 7 tasks with different questions on each task. The learners were allowed to choose topics for speaking randomly with knowing the contents of the tasks. They were given 2 minutes to read and understand the task, and they are allowed to change the task once. After warming-up questions learners were asked to speak about the given tasks and the whole process was made interactive. The tasks were designed from the course book on listening and speaking skills.

Participants

a. Students

The participants of the study were 62 CFY students who study English language skills for the first two semesters as a prerequisite for entering into their majors. The age of the participants was 17-19 years. All the participants were male learners. The test was taken after the completion of the first semester. All the students had the same level of English language proficiency as per their entrance test.

b. Raters

Six raters were involved in the scoring procedure of the speaking test. All the raters are administering this kind of test since 2012. They also had the training sessions for the speaking test. They are regular staff members of CFY program and have a master's degree in English and CELTA teaching certificate. The age of the raters was between 32 to 50 years. The rating procedure was done in pairs, one student and two raters and average scores were awarded to the students.

Instruments

A speaking test and students' scores were used to collect the data. The test was designed by the administration. The test contains different tasks developed from the coursebook. Each task needs 7 to 15 minutes for completion. The total score of the test was 15 for three components of the speaking test, 5 scores of each including task completion, fluency and accuracy, and vocabulary usage. Raters were provided with the speaking criteria, rubrics, and speaking evaluation sheet for each student.

Data analysis

Kuder Richardson's correlation statistical measures are generally involved to estimate the reliability of the test. Test/retest, split-half method, and parallel form are administered for the reliability of the test. However, Underhill and Nic (1987) indicate that these traditional methods of reliability estimation have little association for the oral assessment as they are designed for the set numbers of pre-planned questions. The practical information for the speaking test could be reached by making a comparison of raters scores with others and with two different measures. Conferring to this inter-rater reliability was employed to estimate the reliability of the speaking test for the present test. Ranganathan *et al.* (2017) assert that rater reliability can be gained by utilizing correlation, regression, and the Bland Altman test. To the end, two tests; Bland Altman and Correlation were used to calculate the reliability. SPSS 22 was used to analyze the data for both the tests.

Results

This section presents the results of the data analysis of the study. The results are present in two steps to estimate the reliability of the spoken performance test. Students were awarded on the basis of 15 scores on the speaking test from both the raters, who tested them at the same time. In the first stage, the inter-rater reliability of the test was examined. The use of two distant tests to gauge the reliability of the speaking test was because of the human involvement in the test procedure. As this is the examination of productive skill testing, the rater's decisions of allocating scores may affect the speaking performance test. At the outset, inter-rater reliability was calculated by using correlation coefficients in SPSS software on the scores of the participants of all the raters. The raters were divided into 3 pairs to calculate the correlation. The table below presents the inter-rater reliability.

Table 1: Correlation of the raters' in pairs

Pair 1		Rater A	Rater B
Rater A	Pearson correlation	1	.710(**)
	Sig. (2-tailed)		0.00
	N	62	62
Rater B	Pearson correlation	.710(**)	
	Sig. (2-tailed)	0.00	
	N	62	
Pair 2		Rater C	Rater D
Rater C	Pearson correlation	1	.690(**)
	Sig. (2-tailed)		0.00
	N	62	62
Rater D			
	Pearson correlation	.690(**)	
	Sig. (2-tailed)	0.00	
	N	62	
Pair 3		Rater E	Rater F

Rater E	Pearson correlation	1	.640(**)
	Sig.(2-tailed)		0.00
Rater f	N	62	62
	Pearson correlation	.6400(**)	
	Sig.(2-tailed)	0.00	
	N	62	

The inter-rater reliabilities of the 6 raters are presented in 3 pairs in the table above. The correlation coefficient was assumed for each of the three pairs. The data analysis was estimated by pairing 6 raters in 3 pairs. The correlation of the raters' scores was measured at 0.710, 0.690, and 0.640 respectively, for three pairs. The correlation of the first pair was 0.710, second was correlated .690, and 3rd pair was .640. The reliability of the first pair is acceptable, whereas the reliability of the second and third pair was fairly low. Despite the low reliability of the second and third pair, the p-value of all the pairs was less than measured (0.00) which is smaller than (0.05) which is quite significant.

Reliability on Bland-Altman Test

Bland-Altman test is used to see the agreement between the raters. The raters' scores were paired in three groups to estimate the agreement between raters to see the inter-rater reliability. The figure below illustrates the agreement between Pair A and B

.

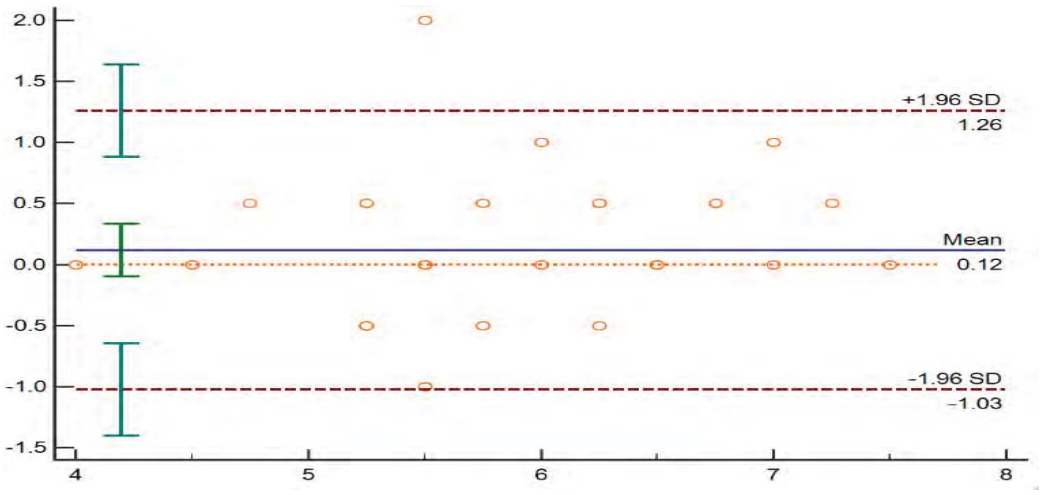

Figure 1: Mean of Pair A & B

The agreement between rater A and B is present in figure 1. It is obvious from figure 1 that most of the points lay near to the mean value and zero, which is an indication of the agreement between the raters. If more than 50% of the points lie near to zero, this shows the agreement between the raters. Moreover, the mean value of the Pair A and Pair B is also near to +1.96 SD and -1.96. The value of the SD of the pair A and pair B is 1.26 and -1.03 which are well in the norm of data to show the agreement. Figure 2 displays the agreement of the raters' scores of pair C and pair D.

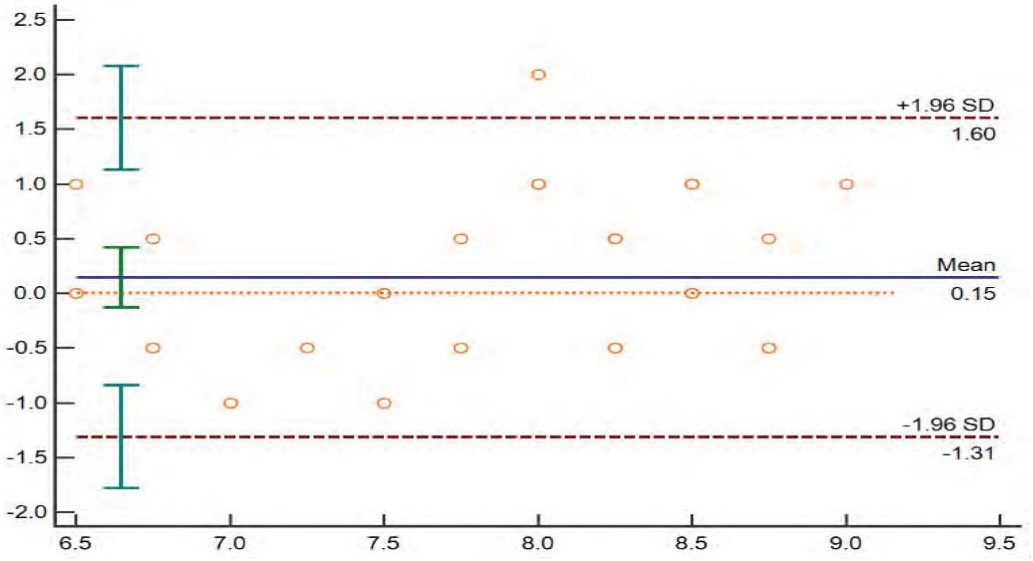

Figure 2: Mean of Pair C & D

The agreement between rater C and D is shown in figure 2. It is also noticeable from the figure that most of the points lie near to the mean value and zero line, which is an indication of the agreement between the raters. Moreover, the mean value of the Pair C and Pair D is also near to +1.96 SD and -1.96. The value of the SD of the pair A and pair B is 1.60 and -1.31which are well in the norm of data to show the agreement. Figure 3 illustrates the rater agreement of pair E and F.

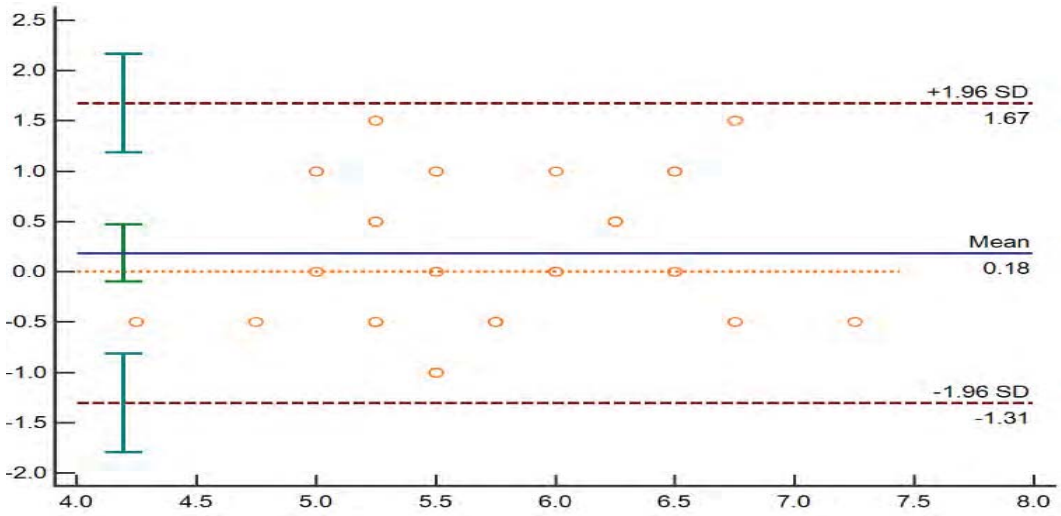

Figure 3: Mean of Pair E & F

The agreement between rater E and F is revealed in figure 3. It is apparent from the figure that most of the points lie near to the mean value and zero line, which is an indication of the agreement between the raters. If more than 50% of the points lie near to zero, this shows the agreement between the raters. Moreover, the mean value of the Pair E and Pair F is also near to +1.96 SD and -1.96. The value of the SD of the pair E and pair F is 1.67 and -1.31 which are well in the norm of data to show the agreement.

Discussion and Recommendation

The reliability of the speaking test was investigated in two ways. The results of the correlation coefficient indicated that the inter-rater reliability of the rater is not satisfactory to meet the desired norm of the test reliability. However, the reliability of the first pair was .710 which is considered satisfactory. The reliability of the second and third pairs is valued at .690 and .640 which is questionable. Although, the reliabilities of the pairs did not seem satisfactory yet the p-values of all three pairs were significant which are less than p= 0.00 which is less than 0.05. This states that the speaking test used at CFY is reliable. The discrepancy in the findings of the inter-rater reliabilities maybe because of the reason the correlation determines that how many same scores were

awarded to the participants, which is not possible where the scores are awarded in point and above than zero.

This leads to the administration of the Bland-Altman test which shows the agreement between two raters. The results of the Bland-Altman indicated that all three pairs of raters showed agreement. The points of the data are closer to the zero line. If more than 50% of the points lie near to zero, this shows the agreement between the raters. This was obvious in all three pairs. Moreover, the mean value of Bland-Altman was also close to +1.96 and -1.96 in all three figures. To this end, it can be interrupted that the speaking test used at CFY is reliable. The assessment of the reliability of the speaking proficiency is not an easy task and it is affected by many aspects including the construct of the test, the task of the test, the knowledge of the learners' background.

The finding of the study is partially aligned with the finding of Fujinaga *et al.* (2007) who indicate that the results of such analysis offer significant value in explaining the reliability of the test. The result of the present attempt was also found in partial agreement with Ozer *et al.* (2014) who carried out the reliability of an oral test. The results of the study asserted that the speaking test was highly reliable, however, the reliability of the present study appeared to the accepted norm of the reliability. This may be a result of the rating criteria and raters have awarded scores in points too, which led to the lower level of the reliability.

The results are consistent with Restrepo and Villa (2003) study which showed inconsistencies in the scores awarded by the examiners. The inconsistencies in the raters' scores have mainly resulted as one of the raters have award hikers scores in grammar and vocabulary use. This can be improved by giving training to the raters. Likewise, the findings are partially aligned with Iwashita *et al.* (2008) who studied the form of speaking proficiency tests in order to establish a rating scale for ESP. The findings implied that certain features of the test had a resilient influence on the total scores assigned by the evaluators which include fluency and vocabulary. Finally, the results also endorsed the result of several investigations (Karim & Haq, 2014; Li, 2019; Quaid, 2018; Read & Nation, 2006) which determined the reliability of the IELTS speaking test. The studies indicated that most of the IELTS speaking test is primarily valid and reliable.

Conclusion

The results of the study indicate that speaking test used for CFY students was moderately reliable. The study can be developed in many ways. To begin with the number of raters, the participant can be increased, and paired can be exchanged for the scoring purposes. The rater training before the test-taking can also present different results. The rater reliability of the speaking performance showed some adverse variation between the shores of the rates. It would be handy, if the grading procedure made clearer to the rater, which could help in making the test more reliable.

References

Alrefaee, Y., Mudkanna, A., & Almansoob, N. T. (2020). Refusals of Suggestions and Offers: An Interlanguage Pragmatic Study. *The Asian ESP Journal*, 16-2-1, 176-195.

Bachman, L. F., Carr, N., Kamei, G., Kim, M., Pan, M. J., Salvador, C., & Sawaki, Y. (2002). A reliable approach to automatic assessment of short answer free responses. *Proceedings of the 19th international conference on Computational linguistics-Volume 2*, pp. 1-4.

Bahrani, T., & Soltani, R. (2011). Improving the components of speaking proficiency. *Canadian Social Science,* 7(3), 78-82.

Bautista, J., Samonte, I., Improgo, C. M., & Gutierrez, M. R. (2020). Mother Tongue versus English as a Second Language in Mathematical Word Problems: Implications to Language Policy Development in the Philippines . *International Journal of Language and Literary Studies, 2(2),* 18-29. https://doi.org/10.36892/ijlls.v2i2.283.

Bygate, M. (2009). 23 Teaching and Testing Speaking. *The handbook of language teaching*, 412.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28(2),* 201-219.

Clark, J. L., & Swinton, S. S. (1979). An exploration of speaking proficiency measures in the TOEFL context. *ETS Research Report Series,* 1979(1), i-69.

Dimitrova-Galaczi, E. (1969). Peer-peer interaction in a paired speaking test: The case of the First Certificate in English.

Franklin, C., Ballan, M., & Thyer, B. (2001). Reliability and validity in qualitative research. *The Handbook of Social Work Research Methods*, 2, 273-292.

Fujinaga, C. I., Zamberlan, N. E., Rodarte, M., & Scochi, C. (2007). Reliability of an instrument to assess the readiness of preterm infants for oral feeding. *Pro-fono: revista de atualizacao cientifica,* 19(2), 143-150.

Fulcher, G. (2013). *Practical Language Testing*: Routledge.

Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing*, 20(3), 321-344.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The Qualitative Report*, 8(4), 597-607.

Gwet, K. L. (2014). *Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters*. Advanced Analytics, LLC.

Iwashita, N., Brown, A., & McNamara, T. (2008). *Assessed levels of second language speaking proficiency: How distinct Applied linguistics*.

Karim, S., & Haq, N. (2014). An Assessment of IELTS Speaking Test. *International Journal of Evaluation and Research in Education,* 3(3), 152-157.

Kemiläinen, E. (2018). *Teaching and assessing oral skills in the advent of oral language testing in the Finnish Matriculation Examination*.Thesis. Master's thesis. University of Helsinki.

Kernot, J., Olds, T., Lewis, L. K., & Maher, C. (2015). Test-retest reliability of the English version of the Edinburgh Postnatal Depression Scale. *Archives of Women's Mental Health,* 18(2), 255-257.

Latief, M. A. (2016). Reliability of Language Skills Assessment Results. *Jurnal Ilmu Pendidikan,* 8(3).

Li, J. (2019). An Evaluation of IELTS Speaking Test. *Open Access Library Journal, 6*(12), 1-17.

Li, W. (2011). Validity Considerations in Designing an Oral Test. *Journal of Language Teaching and Research, 2*(1), 267.

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*(4), 479-499.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing,* 26(3), 397-421.

May, L. (2011). *Interaction in a paired speaking test: The rater's perspective* (Vol. 24): Peter Lang.

McCarthy, M., & O'Keeffe, A. (2004). 2. Research in the Teaching of Speaking. *Annual Review of Applied Linguistic*s, 24, 26-43.

Namaziandost, E., & Nasri, M. (2019). The impact of social media on EFL learners' speaking skill: a survey study involving EFL teachers and students. *Journal of Applied Linguistics and Language Research,* 6(3), 199-215.

Ockey, G. J. (2018). Oral language proficiency tests. *The TESOL Encyclopedia of English Language Teaching*, 1-5.

Ozer, I., Fitzgerald, S. M., Sulbaran, E., & Garvey, D. (2014). Reliability and content validity of an English as a foreign language (EFL) grade-level test for Turkish primary grade students. *Procd Soc Behv, 112*, 924-929.

Quaid, E. D. (2018). Reviewing the IELTS speaking test in East Asia: theoretical and practice-based insights. *Language Testing in Asia, 8*(1), 2.

Ranganathan, P., Pramesh, C., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Measures of agreement. *Perspectives in Clinical Research,* 8(4), 187-191.

Read, J., & Nation, P. (2006). An investigation of the lexical dimension of the IELTS speaking test. *International English Language Testing System (IELTS) Research Reports 2006:* Volume 6, 1.

Restrepo, A. P. M., & Villa, M. E. Á. (2003). Estimating the validity. *REVISTA Universidad EAFIT, 39*(132), 65-75.

Rohan-Minjares, F., Schutzman, E. Z., Chavez, M., Galicia, R., & Valverde, C. (2019). Oral Proficiency Language Testing for Medical Students.

RSari, N. K., & Nike, S. (2017). The use of oral language assessment in learning speaking in junior high school.

Shariq, M. (2020). Feedback and Speaking Skills in Task-Based Language Teaching: Proposed Corrective Measures for EFL Learners. *Asian ESP Journal*. 16(2), 232-248.

Silvia, P. J. (2011). Subjective scoring of divergent thinking: Examining the reliability of unusual uses, instances, and consequences tasks. *Thinking Skills and Creativity,* 6(1), 24-30.

Steenson, J., Vivanti, A., & Isenring, E. (2013). Inter-rater reliability of the Subjective Global Assessment: a systematic literature review. *Nutrition,* 29(1), 350-352.

Tuan, L. T. (2012). Teaching and assessing speaking performance through analytic scoring approach. *Theory and Practice in Language Studies,* 2(4), 673.

Underhill, N., & Nic, U. (1987). *Testing Spoken Language: A Handbook of Oral Testing Techniques*. Cambridge University Press.

Vera, J. D., & Vera, P. D.(2018). Oral Communication Skills in English among Grade 11 Humanities and Social Sciences (HUMSS) Students. *Asian ESP Journal*, 14 (5), 30-52.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing,* 30(2), 231-252.