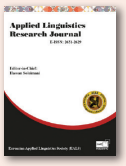
Applied Linguistics Research Journal

ALR Journal 2019;3(2):70–82

Original Research

doi: 10.14744/alrj.2019.83792

The Assessment of Oral Proficiency through Holistic and Analytic Techniques of Scoring: A Comparative Study

Ehsan Namaziandost ,[1]
Sheida Ahmadi,[2]

[1]PhD Student, Department of English, Shahrekord Branch, Islamic Azad University, Shahrekord, *Iran*

[2]State University of Malayer, Hamedan, *Iran*

ABSTRACT

It is an acutely hard and intricate matter to assess the skill of speaking. Holistic and analytic scoring are usually utilized as two methods of testing the speaking performance. In the current study, these two methods of evaluating the spoken proficiency were examined in depth. English speaking skills of a total of 70 subjects, who were Iranian third-grade university EFL learners, were assessed by an interlocutor and an assessor. The interlocutor carried out the holistic scoring while the assessor conducted the analytic scoring. Categories within the analytic scoring comprised of content and organization, pronunciation, vocabulary, and grammar. The analytic average for the four criteria was 3.396, while the holistic scoring mean was 3.628. The findings revealed that there existed a statistically significant difference between the holistic and analytic methods of assessment as the p-value was calculated at 0.002 ($p < 0.05$). It is, thus, suggested that applying both techniques of scoring in the assessment process might be regarded suitable as they seem to supplement each other, and together help towards more comprehensive assessment.

Keywords: Assessing Speaking Skills; Holistic Scoring; Analytic Scoring; Teaching English as a Foreign Language.

Corresponding Author: Ehsan Namaziandost ; PhD Student, Department of English, Shahrekord Branch, Islamic Azad University, Shahrekord, *Iran*

Phone: +98 909 210773832
e-mail: e.namazi75@yahoo.com

Article citation: Namaziandost, E. & Ahmadi, Sh. (2019). The assessment of oral proficiency through holistic and analytic techniques of scoring: A comparative study. Applied Linguistics Research Journal, 3(2): 70–82.

Received Date: 31 March 2019
Accepted Date: 4 April 2019
Online Date: 7 May, 2019

Publisher: Kare Publishing

E-ISSN: 2651-2629

1. Introduction

Among the four skills, the speaking skill is of vital importance (Khamkhien, 2010) as those who have sufficient information about a language are mostly alluded to as speakers of that specific language (Ur, 2012). Analogously, Pokrivčáková (2010) corroborates that numerous foreign language educators and students assume speaking skills as the extent of knowing a language. Göktürk (2016, p. 71) additionally annexes remarkable emphasis to speaking performance: "[w]ith the proliferating significance belonging to speaking as part of one's language ability within the Communicative Language Teaching sample, the instructing of speaking skills in second language learning has become an enthusiastic zone of research over the past two decades". It is likewise the computerized and globalization time which plays an incredible tool since effectual oral communication skills have ascertained to be extremely fundamental in this day and age (Murugaiah, 2016). However, simultaneously, speaking

can be considered as the utmost troublesome skill to need as language has to be produced swiftly and unplanned, which needs a great deal of practice (Anderson, 2015). Indubitably, it takes a lot of time and permanent endeavor for a foreign language learner to become proficient in the speaking skills.

With regard to the assessment of speaking proficiency, O'Sullivan (2012, p. 234) asserts that "[i]t is customarily believed that the most troublesome tests to expand and execute are tests of spoken language ability". In the same vein, Chuang (2009) affirms that since there exist numerous internal and external factors which affect assessors, assessing speaking performance appears to be one of the uttermost hard tasks to administer. Furthermore, Luoma (2004) additionally claims that assessing speaking is challenging because there are numerous agents which affect the perception of an assessor regarding how excellent a person can speak. Moreover, assessors envisage test scores to be precise and suitable for the objectives of appraising spoken proficiency, which is not evermore the case. Therefore, executing appropriate and accurate assessment of speaking performance is a partly hard task, and many aspects require to be considered.

2. Literature Review

2.1. Approaches to assessing speaking skills

Holistic and analytic scoring are two approaches to evaluate the oral proficiency which are commonly utilized for assessment (Al-Amri, 2010; Goh & Burns, 2012; Sarwar, Alam, Hussain, Shah, & Jabeen, 2014; Xi, 2007). The holistic scoring can be additionally alluded to as impressionistic or worldwide scale (Pan, 2016). The holistic approach is concerned with giving an overall score, considering the performance as a whole (Baryla, Shelley & Trainor, 2012; Griffith & Lim, 2012; Helvoort, 2010; Reddy, 2007; Schunn, Godley & DeMartino 2016). "An analytic or profile approach, on the other hand, tries to segregate out notable properties of execution and to assess every one exclusively and freely on its own subscale; the analytic approach thus therefore concentrates consideration on discrete characteristics of execution, normally mixing scores on the detached subscales to generate an overall score for speaking, and sometimes reporting the sub-scores too to give a more extravagant and wealthy dimension of source information, which can be beneficial for diagnostic objectives to manage future instructing/learning goals" (Taylor & Galaczi, 2011, p. 177). Consequently, a few particular criteria are utilized within analytic rubrics (Allen & Tanner, 2006; Babik, Gehringer, Kidd, Pramudianto, & Tinapple, 2016). It is evident that the holistic way of scoring is less time-consuming and less intricated than the analytical approach. However, the analytical way of scoring gives abundant information on the language capability of a learner (Kondo-Brown, 2002). In addition, the rating precision is expanded as raters' consideration is attracted to the particular criteria of language execution (Luoma, 2004). In spite of the way that worldwide and analytical methods for scoring vary from the theoretical viewpoint, they constantly overlap to some extent (Taylor & Galaczi, 2011).

2.2. Analytic scoring

In testing speaking, analytic approach investigates different characteristics of the test independently, scoring each property severalty (Richards & Schmidt, 2013). Utilizing analytical way of scoring within the assessment of spoken execution produces various advantages. Tuan (2012) states that it propounds effective diagnostic information on an examinee's speaking ability, yielding ample perspicacity into the strengths and weaknesses of a learner. Jonsson and Svingby (2007) mention that it is additionally the stability of scoring amongst learners, assignments, and various raters which is expanded. Moreover, utilizing analytical scoring boost the reliability of assessment (Dogan & Uluman, 2017; Kaba & Sengül, 2016). Eventually, Finson, Ormsbee, and Jensen (2011, p. 181) claim that "[a]nalytic rubrics bolster a progressively objective and reliable evaluation of learner work". Expanded objectivity and consistency really emerge out of utilizing the assessment of a few highlights of spoken test. Although the analytic method to the assessment of spoken proficiency presents various considerable advantages, it likewise has a few drawbacks. It is more time-consuming and wearisome because assessors require to give distinct scores for various

parts of a candidate's execution (Aleksandrzak, 2011; Saritha, 2016; Shatrova, Mullings, Blažejová, & Üstünel, 2017). Besides, examiners have to be instructed in order to reliably distinguish among different aspects and components of execution regarding how they are characterized in the rubrics (Vafaee & Yaghmaeyan, 2015). Another detriment is halo effect which refers to the rating within one scale can affect the rating on another scale (Myford & Wolfe, 2003). Finally, Llach (2011, p. 57) indicates that "[o]ne of the major drawbacks of analytic scoring is the hardness in giving obvious and unequivocal definitions for each descriptor". However, despite the fact that analytical scoring has some drawbacks, its merits appear to outweigh and preponderate the disadvantages, and adopting this way of scoring within the assessment of speaking performance can be regarded impartially suitable.

2.3. Analytic scoring criteria

As far as the concrete divisions within analytic rubrics are concerned, Pan (2016) clarifies that dimensionality for the assessment of spoken proficiency may, say, incorporate fluency, vocabulary, and accuracy. The Council of Europe (2001) comprises the accompanying components of spoken language: range, accuracy, fluency, interaction, and coherence. According to Davies (1999) regularly utilized classes within speaking tests are pronunciation or intelligibility, fluency, accuracy, and appropriateness. on the other hand, Gondová (2014, p. 162) elucidates that "the accompanying criteria are regularly utilized: appropriateness, organization of ideas, fluency, grammatical accuracy and the range of grammatical structures, the range of vocabulary and its accuracy, content, pronunciation and intonation, and interaction" (Metruk, 2018). The analytical assessment scales within Cambridge English First certificate consist of grammar and vocabulary, discourse management, pronunciation, and interactive communication (Cambridge English: Understanding Results Guide, 2014). Tuan (2012) states that "based on the objective of the assessment, speaking performance might be evaluated on such criteria as content, organization, cohesion, register, vocabulary, grammar, or mechanics" (p. 673).

2.4. Amount of criteria

It is flagrant that the choice of specific categories must emerge out of the objective of evaluation. However, assessors should be cautious about the quantity of classifications they utilize when they evaluate speaking. Their amount is typically made between three and seven (Ruammai, 2014). On the other hand, Finson, Ormsbee, and Jensen (2011) express that three to six classes are generally utilized. However, the questions are raised about the maximum number of criteria. "Received wisdom is that more than 4 or 5 classifications begins to cause cognitive overload and that 7 classifications are psychologically an upper bound" (Council of Europe 2001, p. 193). Correspondingly, Green (2014), Razali and Isra (2016), and Thornbury (2005) affirm that four to five criteria seem to be the most noteworthy reasonable number regarding assessing spoken proficiency, while Luoma (2004) views five to six categories to be the maximum. It appears to be sensible to accept that it is next to inconceivable for assessors t to concentrate on higher measure of criteria than five or six, and lead reasonable and reliable evaluation in the meantime. "However, prior researches have not given sufficient experimental proofs to help the designation of ideal number of criteria inside rating scales" (Chen, 2016, p. 52).

2.5. Research background

The relationship between the holistic and analytic scoring of Iranian university EFL students' English spoken proficiency was investigated in this study. The participants – the third-grade students of the study program Teaching English Language participated six semesters of the English Language course, which was instructed based the principles of Communicative Language Teaching (CLT). At the end of their last semester, the learners took an speaking exam at the C1 level according to the Common European Framework of Reference for Languages (CEFR) in the form of an interview between an interlocutor and a candidate. Both types of scoring, the holistic and analytic types, were applied. The holistic scoring was executed by the interlocutor, while its analytic

counterpart was carried out by an assessor.

Content and organization, pronunciation, vocabulary, and grammar were the four analytic criteria. The participants could attain the minimum of one and the maximum of five points within every category based on the descriptors for every point, which computed for the absolute of 20 points.

The content and organization category contained the relevancy of replies to items, proper creation of short and long utterances, and responding the items so that the communicative goal was done.

The primary focal point of the pronunciation part was coordinated towards comprehensibility alongside the best possible enunciation of individual phonemes and suitable utilization of stress and intonation. Because of the fact that L2 speakers' English utterances typically show freak phonetic recognitions with regard to their L1 (Bilá, 2010), negligible and trivial properties of L1 (Iranian) accent in the participants' production were not penalized.

The grammar and vocabulary criteria estimated range, as well as precision. As far as the vocabulary category as such is involved, Topkaraoğlu and Dilman (2014) demonstrate that the number of words an L2 learner knows does not appear to be adequate; the participants additionally require to have considerable amount of information about the words they have procured if they desire to become effectual and efficient users of a foreign language. Finally, attention was also devoted to grammar. Similar to vocabulary, both grammatical range and accuracy were inspected.

Regrading the holistic scale, the learners could attain the minimum of one and the maximum of five points based on the descriptors for each point. Thus, the participants were able to reach a sum of 25 points for the entire assessment (holistic scoring + analytic scoring). For instance, a learner reached 4 points for content and organization, 3 points for pronunciation, 4 points for vocabulary, and 3 points for grammar from the assessor, and the interlocutor gave them 3 points. Altogether, they scored (4 + 3 + 4 + 3 + 3) 17 points out of 25, which establishes 68%. Thereafter, the participants were given a score based on the university scoring scales criteria.

The following research questions were formulated.

1. What scores do the subjects achieve in the four categories of analytic scoring?
2. What is the average score with regard to the holistic scoring?
3. What is the average score with regard to the analytic scoring?
4. What is the difference between the holistic and analytic scoring? Is the difference statistically significant?

3. Method

3.1. Participants

The participants of this study were 70 third-year university EFL students who were studying English Language Teaching at Islamic Azad University of Abadan, Iran. The participants in this research were both male (n=40) and female (n=30) with the age range from 23 to 27. The interlocutor and assessor were two Iranian university professors holding Ph.D. in TEFL. They had almost seven years of experience in the assessment of spoken proficiency when the assessment was executed, and the assessor had completed two semesters of assessing English language course as a part of his master and Ph.D. studies.

3.2. Instrument and procedures

The participants were randomly assigned a topic on which they were needed to hold an interview with the interlocutor. They were not given any time for preparation. The interlocutor asked opinion-based open questions, which were within the range of general knowledge of the subjects, so that the assessment process was not negatively affected by testing knowledge rather than speaking skills. The assessor was not present so as not to distract or impact the participants in any way. He

was taking notes in order to make his assessment as reliable as possible. The examination lasted approximately 15 minutes. Afterwards, a participant was asked to wait outside the room, so that the interlocutor and assessor could award points to the participant for their performance. When the total mark was computed, the participant was called back to the room to dispute how they did in the speaking test. Each participant was given a beneficial feedback on how they performed within every category.

3.3. Results

The analytic scoring marks with the scores for each category (content and organization, pronunciation, vocabulary, grammar) are illustrated in Table 1 (see Appendix A). The table also includes the mean values for all the subjects' performance in the four categories. The data demonstrate that the participants were most successful in the category content and organization (3.928), and attained the bottommost scores in the grammar category (2.514). The pronunciation and vocabulary parts indicate the scores of 3.471 and 3.271 respectively.

Among the four sections, the content and organization section were the minimum problematical ones. The participants were remunerated for not sticking to the point, or when the questions were not replied, and utterances were either unrelated, not fluent, or of an inopportune length. The pronunciation category included both segmental and suprasegmental fallacies. The sections frequently involved the replacement of English phonemes, especially those which do not exist in subjects' L1, for Iranian sounds. "Both teachers and learners require to reminisce that substituting some sounds for others hampers communication and mostly causes a menace to intelligibility" (Metruk, 2017, p. 15). The highest frequent mistake within the prosodic features was the word stress. Regarding the vocabulary and grammar categories, the learners faced remarkable laboriousness and problems with the range of lexis, and experienced even outstanding troubles with the range of grammatical structures.

Table 2 (see Appendix B) indicates the average analytic scoring mark for each participant. For instance, if a participant got 5 points for content and organization, 3 for pronunciation, 4 for vocabulary, and 3 for grammar, the average mark for the analytic scoring is 3.5 (5 + 3 + 4 + 3 = 15, and this number was divided by the number of categories: 15 ÷ 4 = 3.75). The mean for the holistic scoring for all the participants, which is additionally illustrated in Table 2 (Appendix B), was 3.628, while the average value of analytical scoring for all the participants was 3.396. Although the difference between the holistic and analytic scoring is only 0.232 (3.628– 3.396= 0.232), the p-value which stands for the level of statistical significance was calculated at 0.002, which means that there is a statistically significant difference between the analytic and holistic scoring ($p < 0.05$). Therefore, the study outcomes reveal that the analytic method of scoring proved somewhat more exact and reliable way of assessing the spoken proficiency than the holistic approach. Moreover, the participants were provided with meticulous feedback on how successful they were in each category as the assessor took notes during the exam. The analytic scoring also presented diagnostic information so that the teachers understood which areas the EFL learners require to pay excessive attention to in the future.

4. Discussion and Conclusion

This study tried to investigate the holistic and analytic way of assessing speaking skills in a higher-education context. Totally 70 third-year university students undertook an speaking exam in the TEFL at Abadan Islamic Azad University in Iran. The examination was at the C1 level based on CEFR. Both holistic and analytic ways of scoring were utilized.

The findings indicate that the participants in the four categories – content and organization, pronunciation, vocabulary, and grammar, obtained 3.928, 3.471, 3.271 and 2.514 points respectively. In spite of the fact that CLT should be the essential method of TEFL, it seems that L2 learners encounter difficulties when they need to utilize B2/C1 level words along with more intricate and dummy grammatical structures within their utterances. This may be the outcome of utilizing the Grammar Translation Method (in its different structures) to some degree in Iranian

system of education. The students may even know the B2/C1 words; however, they are not ready to utilize them when they speak. Thus, following the principles of CLT, and providing EFL learners with sufficient space for practicing speaking might prove useful.

Moreover, the outcomes uncover that the average score for holistic and analytic ways of scoring was 3.628 and 3.396, respectively. The p value was computed at 0.002; therefore, we found a statistically significant difference between the holistic and analytic methods of scoring ($p < 0.05$). This does not really imply that one method of scoring is more reliable than the other as the subjectivity of the assessor and interlocutor may have played its role. However, applying both ways of scoring in the assessment process can be viewed as valuable and proper since the two methods appear to supplement each other. Additionally, the analytic scoring enabled the subjects to be furnished with a rather detailed feedback on their performance in specific categories. At long last, the results propounded beneficial diagnostic information so that both the EFL higher-education students and their teachers know which areas they should focus on to a greater extent.

This study suffered from several limitations. First, there was only one assessor (and one interlocutor) and their subjective perception and interpretation of a participants' speaking performance might have affected the assessment process. However, it should be mentioned that the evaluation of spoken proficiency is an extremely subjective process, and there are innumerable factors which affect the assessor's judgement (Jankowska & Zielińska, 2015). It is therefore proposed that future studies utilize an excessive number of assessors in order to give enough statistical power for the evaluation of the relationship between the holistic and analytic ways of scoring. In the same way, a monumental sample of participants can be participated in future researches as well. Moreover, the description of bands in the analytical scoring scales may have also played its part within the assessment process. In addition, a subjective interpretation might have affected the assessment process. Nevertheless, it should be accentuated that it is a rather hard task to propose clear-cut and unequivocal definitions for the descriptors (Llach, 2011). It appears sensible to presume that the level of subjectivity can be lessened by undergoing a suitable instruction and by obtaining years of experience, so that the assessment can become as precise, reliable, and objective as possible. Ultimately, comparing the difference between female and male scores within the assessment of spoken proficiency in future studies might be engrossing.

It can be deduced that synthesizing the analytic and holistic scoring may be considered as a rather viable choice when it comes to the assessment of speaking skills. Both ways have their merits and drawbacks, and utilizing these two methods of scoring might probably lead to a more objective scoring.

References

Al-Amri, M. (2010). Direct spoken English testing is still a real challenge to be worth bothering about. *English Language Teaching*, 3 (1), 113-117.

https://doi.org/10.5539/elt.v3n1p113

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

Allen, D. & Tanner, K. (2006). Rubrics: Tools for making learning goals and evaluation criteria explicit for both teachers and learners. *Life Sciences Education*, 5 (3), 197–203.

http://doi.org/10.1187/cbe.06-06-0168

Anderson, J. (2015). *A guide to the practice of English language teaching for teachers and trainee teachers*. Nairobi: East African Educational Publishers Ltd.

Babik, D., Gehringer, E., Kidd, J., Pramudianto, F. & Tinapple, D. (2016). Probing the landscape: Toward a systematic taxonomy of online peer assessment systems in education. Paper presented at the CSPRED 2016: *Workshop on Computer-Supported Peer Review in Education*, Raleigh, NC

Baryla, E., Shelley, G., & Trainor, W. (2012). Transforming rubrics using factor analysis: Practical assessment. *Practical Assessment Research & Evaluation*, 17 (4).

Bilá, M. (2010). Perception and production of a second language and the concept of a foreign accent. In S. Pokrivčáková et al. (Eds.) *Modernization of teaching foreign languages: CLIL, inclusive and intercultural education*, pp. 123-143. Brno: Masaryk University.

Cambridge English: Understanding Results Guide. (2014). http://www.gml.cz/prof/zajickova/Cambridge%20exams_information/Understanding%20results%20guide.pdf

Chen, G. (2016). Developing a Model of analytic rating scales to assess college students" l2 Chinese oral performance. *International Journal of Language Testing*, 6 (2), 50-71.

Chuang, Y. (2009). Foreign language speaking assessment: Taiwanese college English teachers" scoring performance in the holistic and analytic rating methods. *The Asian EFL Journal*, 11 (1), 150-173.

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

Davies, A. (1999). *Dictionary of language testing*. Cambridge: Cambridge University Press.

Dogan, C. & Uluman. M. (2017). A comparison of rubrics and graded category rating scales with various methods regarding raters' reliability. *Educational Sciences: Theory & Practice*, 17 (2), 631-651. doi:10.12738/estp.2017.2.0321

https://doi.org/10.12738/estp.2017.2.0321

Finson, K., Ormsbee, C., & Jensen, M. (2011). *Differentiating science instruction and assessment for learners with special needs, k-8*. Thousand Oaks: Corwin, A SAGE Company.

https://doi.org/10.4135/9781483387529

Goh, C. & Burns, A. (2012). *Teaching speaking: A holistic approach*. New York: Cambridge University Press.

Gondová, D. (2014). *Taking first steps in teaching English: Assessing learners*. Žilina: EDIS.

Göktürk, N. (2016). Examining the Effectiveness of Digital Video recordings on Oral Performance of EFL Learners. *Teaching English with Technology*, 16 (2), 71-96.

Green, A. (2014). *Exploring language assessment and testing: Language in action*. New York: Routledge.

https://doi.org/10.4324/9781315889627

Griffith, W. & Lim, H. (2012). Performance-based assessment: rubrics, web 2.0 tools and language competencies. *Mextesol Journal*, 36 (1).

Helvoort, J. (2010). A scoring rubric for performance assessment of information literacy in Dutch higher education. *Journal of Information Literacy*, 4 (1), 22-39.

http://dx.doi.org/10.11645/4.1.1256

Jankowska, A. & Zielińska, U. (2015). Designing a self-assessment instrument for developing the speaking skill at the advanced level. In M. Pawlak and E. WaniekKlimczak (Eds.) *Issues in teaching, learning and testing speaking in a second language*. Heidelberg: Springer.

https://doi.org/10.1007/978-3-642-38339-7_16

Jonnson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences.

Educational Research Review, 2 (2), 130-144.
https://doi.org/10.1016/j.edurev.2007.05.002

Kaba, Y. & Sengül, S. (2016). Developing the rubric for evaluation problem posing (REPP). *International Online Journal of Educational Sciences*, 8 (1), 8-25.
https://doi.org/10.15345/iojes.2016.01.002

Khamkhien, A. (2010). Teaching English speaking and English-speaking tests in the Thai context: a reflection from Thai perspective. *English Language Teaching,* 3 (1), 184-190.
https://doi.org/10.5539/elt.v3n1p184

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19 (1), 3-31.
https://doi.org/10.1191/0265532202lt218oa.

Llach, M. (2011). *Lexical errors and accuracy in foreign language writing*. New York: Multilingual Matters.
https://doi.org/10.21832/9781847694188

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
https://doi.org/10.1017/CBO9780511733017

Metruk, R. (2017). Pronunciation of English dental fricatives by Slovak University EFL Students. *International Journal of English Linguistics*, 7 (3), 11-16. doi:10.5539/ ijel.v7n3p11
https://doi.org/10.5539/ijel.v7n3p11

Murugaiah, P. (2016). Pecha Kucha style PowerPoint presentation: An innovative CALL approach to developing oral presentation skills of tertiary students. *Teaching English with Technology*, 16 (1), 88-104.

Myford, C. & Wolfe, E. (2003). Detecting and measuring rater effects using many facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4 (4), 386-422.

O' Sullivan, B. (2012). Assessing speaking. In C. Coombe, P. Davidson, B. O' Sullivan, and S. Stoynoff (Eds.) *the Cambridge guide to second language assessmen*t, pp. 234- 246. New York: Cambridge University Press.

Pan, M. (2016). *Nonverbal delivery in speaking assessment. from an argument to a rating scale formulation and validation*. Singapore: Springer.
https://doi.org/10.1007/978-981-10-0170-3

Pokrivčáková, S. (2010). *Modern teacher of English*. Nitra: ASPA.

Razali, K. & Isra, M. (2016). Male and female teachers roles in assessment of speaking skill: Gender equality: *International Journal of Child and Gender Studies*, 2 (1), 1-10.

Reddy, Y. (2007). Effect of rubrics on enhancement of student learning. *Educate*, 7 (1), 3-17.

Richards, J. & Schmidt, R. (2013). *Longman dictionary of language teaching and applied linguistics* (4th ed.). New York: Routledge.
https://doi.org/10.4324/9781315833835

Ruammai, P. (2014). Constructing scoring instrument for writing assessment and fostering critical thinking. In H. Lee (Ed.) *The international conference on language and communication innovative inquiries and emerging paradigms in language, media and communication*, pp. 127-137.

Saritha, K. (2016). Rubric for English language teaching research. *Research Journal of English Language and Literature*, 4 (2), 725-731.

Sarwar, M., Alam, M., Hussain, S., Shah, A., & Jabeen, M. (2014). Assessing English speaking skills of prospective teachers at entry and graduation level in teacher education program. *Language Testing in Asia*, 4 (5). https://doi.org/10.1186/2229-0443-4-5

Schunn, C., Godley, A., & DeMartino, S. (2016). The reliability and validity of peer review of writing in high school English classes. *Journal of Adolescent & Adult Literacy*, 60 (1), 13-23. doi:10.1002/jaal.525
https://doi.org/10.1002/jaal.525

Shatrova, Z., Mullings, R., Blažejová, S., & Üstünel, E. (2017). English speaking assessment: developing a speaking test for students in a preparatory school. *International Journal of English Language Teaching*, 5 (3), 27-40.

Taylor, L. & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.), M. Milanovic & C. Weir (Series Eds.) *Studies in language testing 30. examining speaking. research and practice in assessing second language speaking*, pp. 171-233. Cambridge: Cambridge University Press.

Thornbury, S. (2005). *How to teach speaking*. Harlow: Pearson Education Limited.

Topkaraoğlu, M. & Dilman, H. (2014). Effects of studying vocabulary enhancement activities on students"vocabulary production levels. *Procedia - Social and Behavioral Sciences*, 152, 931-936. https://doi.org/10.1016/j.sbspro.2014.09.345

Tuan, L. (2012). Teaching and assessing speaking performance through analytic scoring approach. *Theory and Practice in Language Studies*, 2 (4), 673-679. doi:10.4304/tpls.2.4.673-679. https://doi.org/10.4304/tpls.2.4.673-679

Ur, P. (2012). *A course in English language teaching*. Cambridge: Cambridge University Press.

Vafee, P. & Yaghmaeyan, B. (2015). Providing evidence for the generalizability of a speaking placement test scores. *Iranian Journal of Language Testing*, 5 (2), 78-95.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® academic speaking test for operational use. *Language Testing*, 24 (2), 251-286. https://doi.org/10.1177/ 0265532207076365

.

Appendix A

Analytic scoring marks

Participant	Content	Pronunciation	Vocabulary	Grammar
1	4	3	3	4
2	3	3	3	1
3	4	3	3	1
4	3	3	3	2
5	2	3	2	2
6	3	2	2	3
7	4	2	1	3
8	3	2	2	3
9	3	5	3	3
10	4	5	3	4
11	4	4	4	4
12	3	2	4	5
13	3	2	4	3
14	3	3	4	3
15	3	3	4	2
16	5	4	2	2
17	3	4	2	2
18	4	4	2	2
19	5	2	1	1
20	2	3	3	1
21	5	3	5	1
22	5	5	5	2
23	5	5	3	2
24	5	5	5	2
25	2	5	4	2
26	4	5	4	3
27	5	4	3	3
28	3	4	3	3
29	5	4	4	3
30	5	4	2	4
31	4	2	4	5
32	4	2	4	3
33	5	2	4	3
34	5	3	4	2
35	5	3	3	2

36	5	4	4	2
37	4	4	3	2
38	5	4	4	4
39	4	4	3	4
40	4	3	4	3
41	5	4	3	3
42	5	5	5	1
43	3	4	3	2
44	4	4	5	3
45	4	3	2	1
46	5	4	3	4
47	5	3	3	5
48	5	4	1	2
49	5	2	1	2
50	5	2	3	2
51	4	4	3	3
52	4	3	5	3
53	2	5	4	4
54	2	5	4	2
55	3	4	4	5
56	3	3	2	3
57	3	5	2	2
58	4	4	3	2
59	5	4	5	2
60	5	3	5	2
61	4	3	3	2
62	4	3	3	1
63	3	3	4	1
64	3	4	4	2
65	2	2	4	3
66	2	4	3	1
67	5	3	3	1
68	5	3	2	1
69	5	3	3	2
70	4	3	4	3
Mean	3.928	3.471	3.271	2.514

Appendix B

Comparison of Holistic and Analytic Scoring

Participant	Holistic scoring	Analytic scoring mean
1	3	2.75
2	4	2.75
3	3	2
4	5	2
5	3	3.5
6	2	2
7	3	2.5
8	5	2.5
9	5	4
10	3	4
11	4	3.75
12	2	2
13	5	2
14	2	3.25
15	2	3
16	4	2.25
17	3	4
18	5	3.75
19	4	2
20	3	3.75
21	4	3
22	5	3.75
23	2	5
24	4	5
25	3	3.75
26	2	5
27	5	4
28	4	3.25
29	3	4
30	3	4
31	3	2

32	2	3.25
33	3	3.25
34	4	3.25
35	4	3
36	5	4
37	3	3.5
38	4	3.75
39	5	4.5
40	5	3
41	4	4.75
42	2	4.75
43	4	4.75
44	2	4.75
45	3	3
46	3	2.25
47	4	3.25
48	3	2.25
49	5	2.75
50	3	2
51	4	4.25
52	3	3
53	5	4.25
54	4	4.25
55	4	4.25
56	2	3.25
57	4	2.5
58	4	4
59	3	4.25
60	4	3
61	3	3.25
62	5	2
63	3	3
64	4	3.75
65	5	3.75
66	4	3.75
67	4	3.75
68	5	3
69	3	3.5
70	5	4.5
Mean	3.628	3.396