THE JOURNAL OF TEACHING ENGLISH FOR SPECIFIC AND ACADEMIC PURPOSES
Vol. 6, N° 1, 2018, pp. 179–189

UDC: (811.111.::371.8):159.9.072(544) https://doi.org/10.22190/JTESAP1801179M

COMPARING HOLISTIC AND ANALYTIC WAYS OF SCORING IN THE ASSESSMENT OF SPEAKING SKILLS

Rastislav Metruk

Faculty of Humanities, University of Žilina, Slovakia
Phone: +421 41/513 6170, E-Mail: rastislav.metruk@gmail.com

Abstract. *Assessing the skill of speaking is an extremely difficult and complex matter. Two methods of testing oral performance are usually applied: holistic and analytic scoring. In the present study, these two ways of evaluating the spoken proficiency are explored in order to examine the relationship between them. English speaking skills of a total of 50 subjects, who are Slovak university EFL (English as a foreign language) students, were assessed by an interlocutor and an assessor. The interlocutor conducted the holistic scoring, while the assessor performed the analytic scoring. Categories within the analytic scoring consisted of content and organisation, pronunciation, vocabulary, and grammar. The overall average for the four criteria was 3.32, while the holistic scoring mean was 3.56. The results demonstrate that there exists a statistically significant difference between the holistic and analytic ways of assessment as the p-value was calculated at 0.001 ($p < 0.05$). It is, therefore, suggested that employing both ways of scoring in the assessment process might be considered appropriate as they appear to complement each other, and together contribute towards more objective assessment.*

Key words*: assessing speaking skills, holistic scoring, analytic scoring, teaching English as a foreign language*

1. INTRODUCTION

The skill of speaking appears to be the most important of the four skills (Khamkhien, 2010) as people who have knowledge about a language are often referred to as speakers of that particular language (Ur 2012). Similarly, Pokrivčáková (2010) asserts that many foreign language teachers and learners deem speaking skills as the measure of knowing a language. Göktürk (2016, p. 71) also attaches considerable significance to oral performance: '[w]ith the increasing importance attached to speaking as part of one's language competence within the Communicative Language Teaching paradigm, the teaching of speaking skills in second language learning has become a burgeoning area of research over the past two decades'. It is also the digital and globalization era which occupies a powerful role since effective oral communication skills have proved to be really necessary in this day and age (Murugaiah, 2016). However, at the same time, speaking can be regarded as the most difficult skill to acquire as language has to be produced quickly and without planning, which requires a lot of practice (Anderson 2015). Undoubtedly, it takes a great deal of time and constant effort for a foreign language learner to master the speaking skills.

Submitted January 31st, 2018, accepted for publication February 22nd, 2018

As far as the assessment of oral proficiency is concerned, O'Sullivan (2012, p. 234) maintains that '[i]t is commonly believed that tests of spoken language ability are the most difficult to develop and administer'. Chuang (2009) maintains that assessing oral performance seems to be one of the most difficult to carry out because there exist many internal and external factors which affect assessors. Luoma (2004) also asserts that assessing speaking is challenging because there are many factors which influence the impression of an assessor in terms of how well a person can speak. Furthermore, assessors expect test scores to be accurate and appropriate for the purposes of evaluating spoken proficiency, which is not always the case. Thus, performing proper and correct assessment of oral performance is a rather difficult task, and plenty of aspects need to be taken into consideration.

2. Literature Review

2.1. Ways of assessing speaking skills

Two ways of evaluating spoken proficiency are normally used for assessment, namely holistic and analytic scoring (Al-Amri, 2010; Goh & Burns, 2012; Sarwar et al., 2014; Xi, 2007). The holistic scoring can be also referred to as impressionistic or global scale (Pan, 2016). The holistic approach is concerned with providing an overall score, taking the performance as a whole into consideration (Baryla, Shelley & Trainor, 2012; Griffith & Lim, 2012; Helvoort, 2010; Reddy, 2007; Schunn, Godley & DeMartino 2016).

"An analytic or profile approach, on the other hand, seeks to separate out salient features of performance and to evaluate each one individually and independently on its own subscale; the analytic approach thus focuses attention on discrete qualities of performance, typically combining scores on the separate subscales to produce an overall score for speaking, and sometimes reporting the subscores as well to provide a richer level of source information, which can be useful for diagnostic purposes to guide future teaching/learning objectives" (Taylor and Galaczi, 2011, p. 177). Therefore, several distinct criteria are used within analytic rubrics (Allen & Tanner, 2006; Babik et al., 2016).

It is apparent that the holistic way of scoring is less time-consuming and less complicated than the analytical approach. However, the analytical way of scoring provides ample information on the language ability of a candidate (Kondo-Brown 2002). Moreover, the rating accuracy is increased as raters' attention is drawn to the specific criteria of language performance (Luoma, 2004). Despite the fact that global and analytical ways of scoring differ from the conceptual standpoint, they invariably overlap to some extent (Taylor & Galaczi, 2011).

2.2. Analytic scoring

Analytic approach in testing speaking examines various features of the test separately, scoring each feature independently (Richards and Schmidt 2013). Employing analytical way of scoring within the evaluation of spoken performance yields a number of benefits. Tuan (2012) maintains that it offers useful diagnostic information on an examinee's speaking ability, providing more insight into the strengths and weaknesses of a candidate. Jonsson and Svingby (2007) note that it is also the consistency of scoring across students, assignments, and different raters which is increased. Furthermore, employing analytical scoring enhances the reliability of assessment (Dogan & Uluman, 2017; Kaba & Sengül, 2016). Finally, Finson, Ormsbee & Jensen (2011, p. 181) maintain that '[a]nalytic rubrics

support a more objective and consistent assessment of student work'. Increased objectivity and consistency actually arise out of employing the evaluation of several features of spoken test.

In spite of the fact that the analytic approach to the evaluation of spoken proficiency offers a number of substantial benefits, it also has some disadvantages. It is more time-consuming since assessors need to give separate scores for different aspects of a candidate's performance (Aleksandrzak, 2011; Saritha, 2016; Shatrova et al., 2017). Furthermore, examiners have to be trained in order to reliably differentiate between various dimensions and components of performance in connection with how they are defined in the rubrics (Vafaee & Yaghmaeyan, 2015). Another drawback is that the rating within one scale can affect the rating on another scale, which may be referred to as halo effect (Myford & Wolfe, 2003). Finally, Llach (2011, p. 57) points out that '[o]ne of the main disadvantages of analytic scoring is the difficulty in providing clear-cut and unambiguous definitions for each descriptor'. However, despite the fact that analytical scoring has some disadvantages, its benefits seem to outweigh the drawbacks, and adopting this way of scoring within the evaluation of oral performance can be considered fairly appropriate.

2.3. Analytic scoring criteria

As far as the concrete categories within analytic rubrics are concerned, M. Pan (2016) explains that dimensions for the assessment of spoken proficiency may, for instance, include fluency, vocabulary, and accuracy. Council of Europe (2001) includes the following aspects of spoken language: range, accuracy, fluency, interaction, and coherence. Davies (1999) states that commonly used categories within speaking tests are pronunciation or intelligibility, fluency, accuracy, and appropriateness. Alternatively, Gondová (2014, p. 162) explains that the following criteria are commonly used: appropriateness, organisation of ideas, fluency, grammatical accuracy and the range of grammatical structures, the range of vocabulary and its accuracy, content, pronunciation and intonation, and interaction. The analytical assessment scales within Cambridge English First certificate include grammar and vocabulary, discourse management, pronunciation, and interactive communication (Cambridge English: Understanding Results Guide, 2014). Tuan (2012, p. 673) maintains that '[d]epending on the purpose of the assessment, speaking performance might be rated on such criteria as content, organisation, cohesion, register, vocabulary, grammar, or mechanics'.

2.4. Amount of criteria

It is apparent that the selection of particular categories ought to arise out of the purpose of evaluation. However, assessors need to be careful about the number of categories they employ when they assess speaking. Their amount is normally created between three and seven (Ruammai, 2014). Alternatively, Finson, Ormsbee, & Jensen (2011) state that three to six categories are commonly applied. However, the questions are raised about the maximum number of criteria. 'Received wisdom is that more than 4 or 5 categories starts to cause cognitive overload and that 7 categories is psychologically an upper limit' (Council of Europe 2001, p. 193). Similarly, Green (2014), Razali & Isra (2016), and Thornbury (2005) assert that four to five criteria appear to be the highest manageable number in terms of assessing spoken proficiency, while Luoma (2004) considers five to six categories to be the maximum. It seems reasonable to assume that it is next to impossible for assessors to focus on higher amount of criteria than five or six, and conduct fair and reliable assessment at

the same time. 'However, previous studies have not provided empirical evidence to support the determination of optimal number of criteria within rating scales' (Chen, 2016, p. 52).

3. Research Background

This paper explores the relationship between the holistic and analytic scoring of English spoken proficiency of Slovak university EFL students at a university in Slovakia. The subjects – the bachelor students of the study programme Teaching English Language and Literature attended six semesters of the English Language course, which was taught according to the principles of Communicative Language Teaching (CLT). At the end of their last semester, the students took an oral examination 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 employed. The holistic scoring was performed by the interlocutor, while its analytic counterpart was conducted by an assessor.

The four analytic criteria were comprised of content and organisation, pronunciation, vocabulary, and grammar. The subjects could achieve the minimum of one and the maximum of five points within each category according to the descriptors for each point, which accounted for the total of 20 points.

The content and organisation category included the relevance of responses to questions, appropriate production of short and long utterances, and answering the questions so that the communicative purpose was accomplished.

The main focus of the pronunciation section was directed towards intelligibility along with the proper articulation of individual phonemes and appropriate use of stress and intonation. Due to the fact that L2 speakers' English utterances normally display deviant phonetic realizations based on their L1 (Bilá, 2010), minor and insignificant traits of L1 (Slovak) accent in the subjects' production were not penalized.

The grammar and vocabulary criteria measured not only range, but also accuracy. As far as the vocabulary category as such is concerned, Topkaraoğlu & Dilman (2014) indicate that the amount of words an L2 learner knows does not seem to suffice; the learners also need to have substantial amount of information about the words they have acquired if they wish to become efficient and effective users of a foreign language. Finally, attention was also devoted to grammar. Similarly to vocabulary, both grammatical range and accuracy were examined.

As far as the holistic scale is concerned, the students could achieve the minimum of one and the maximum of five points according to the descriptors for each point. Therefore, the candidates were able to achieve a total of 25 points for the whole assessment (holistic scoring + analytic scoring). For example, a candidate achieved 3 points for content and organisation, 4 points for pronunciation, 2 points for vocabulary, and 3 points for grammar from the assessor, and the interlocutor gave them 4 points. Altogether, they scored (3 + 4 + 2 + 3 + 4) 16 points out of 25, which constitutes 64%. Subsequently, the candidates were given a grade according to 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?

3.1. Research sample

The subjects were formed by 50 Slovak third-year university EFL students, 46 female speakers and 4 male speakers, with a mean age of 20 years. Their major was Teaching English Language and Literature, and all of them were of Slovak nationality.

The interlocutor and assessor were two university teachers, one female (lecturer) and one male (senior lecturer). The interlocutor had a master's degree in the English language, and the assessor held a Ph.D. degree in the same field. Both the interlocutor and assessor were of Slovak nationality. They had had approximately five years of experience in the assessment of spoken proficiency when the assessment was performed, and the assessor had completed one semester of assessing English language course as a part of his master studies.

3.2. Instrument and procedures

The subjects were randomly assigned a topic on which they were required 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 influenced by testing knowledge rather than speaking skills. The assessor was not seated within the primary field of vision of the candidates so as not to distract or influence the candidates 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 candidate was asked to wait outside the room, so that the interlocutor and assessor could award points to the candidate for their performance. When the total mark was calculated, the candidate was called back to the room to discuss how they did in the oral test. Each candidate was provided with useful feedback on how their performed within every category.

3.3. Research results

The analytic scoring marks with the scores for each category (content and organisation, pronunciation, vocabulary, grammar) are displayed in Table 1. The table also contains the mean values for all the subjects' performance in the four categories. The data illustrate that the subjects were most successful in the category content and organisation (3.96), and achieved the lowest scores in the grammar category (2.54). The pronunciation and vocabulary sections reflect the scores of 3.54 and 3.22 respectively.

The content and organisation section was the least problematic of the four sections. The candidates were penalised for not sticking to the point, or when the questions were not answered, and utterances were either irrelevant, not fluent, or of an inappropriate length. The pronunciation category involved both segmental and suprasegmental errors. The segments frequently included the substitution of English phonemes, particularly those which do not exist in subjects' L1, for Slovak sounds. "Both teachers and learners need to remember that replacing certain sounds for others hinders communication and often poses a threat to intelligibility" (Metruk, 2017, p. 15). The most frequent error within the prosodic features was the word stress. As far as the vocabulary and grammar categories are concerned, the subjects encountered considerable difficulties with the range of lexis, and experienced even greater problems with the range of grammatical structures.

Table 1 Analytic scoring marks

Subject	Content	Pronunciation	Vocabulary	Grammar
1	3	4	4	3
2	4	4	3	3
3	3	3	3	3
4	4	4	3	2
5	3	3	3	1
6	2	4	3	3
7	3	3	2	1
8	4	3	4	2
9	4	3	4	2
10	3	3	2	2
11	4	3	3	2
12	3	3	3	2
13	3	3	2	1
14	3	3	3	2
15	4	3	3	2
16	3	4	4	3
17	5	3	5	4
18	3	3	2	2
19	2	4	1	1
20	5	3	4	4
21	5	4	4	3
22	5	3	4	2
23	5	3	3	2
24	3	3	2	2
25	5	4	3	2
26	4	3	4	2
27	2	3	2	2
28	5	5	3	3
29	5	3	4	3
30	5	4	3	3
31	5	4	3	3
32	5	5	4	4
33	5	5	4	3
34	3	3	2	2
35	4	3	4	4
36	5	4	4	3
37	4	2	1	1
38	4	3	3	2
39	5	3	3	1
40	3	3	2	1
41	4	3	2	2
42	5	5	5	4
43	5	5	5	4
44	5	4	3	3
45	5	5	5	4
46	5	4	5	5
47	5	5	4	4
48	5	4	5	4
49	2	3	2	2
50	2	3	2	2
Mean	3,96	3,54	3,22	2,54

Table 2 shows the average analytic scoring mark for each subject. For example, if a candidate received 4 points for content and organisation, 3 for pronunciation, 4 for vocabulary, and 3 for grammar, the average mark for the analytic scoring is 3.5 (4 + 3 + 4 + 3 = 14, and this number was divided by the amount of categories: 14 ÷ 4 = 3.5). The mean for the holistic scoring for all the subjects, which is also included in Table 2, was 3.56, while the average value of analytical scoring for all the candidates was 3.32. In spite of the fact that the difference between the holistic and analytic scoring is only 0.24 (3.56 – 3.32 = 0.24), the p-value which stands for the level of statistical significance was calculated at 0.001, which means that there is a statistically significant difference between the analytic and holistic scoring ($p < 0.05$). Thus, the research results indicate that the analytic method of scoring proved slightly more accurate and reliable way of assessing the spoken proficiency than the holistic approach. Furthermore, the subjects were provided with rigorous feedback on how successful they were in each category as the assessor took notes during the examination. The analytic scoring also offered diagnostic information so that the university teachers knew which areas the university EFL learners need to pay more attention to in the future.

Table 2 Comparison of holistic and analytic scoring

Subject	Holistic scoring	Analytic scoring mean
1	4	3,5
2	4	3,5
3	4	3
4	3	3,25
5	3	2,5
6	3	3
7	2	2,25
8	3	3,25
9	3	3,25
10	3	2,5
11	5	3
12	3	2,75
13	2	2,25
14	3	2,75
15	3	3
16	4	3,5
17	5	4,25
18	3	2,5
19	2	2
20	5	4
21	4	4
22	4	3,5
23	3	3,25
24	3	2,5
25	4	3,5
26	3	3,25
27	2	2,25
28	4	4
29	3	3,75
30	5	3,75
31	4	3,75
32	5	4,5
33	4	4,25
34	3	2,5
35	4	3,75
36	4	4
37	2	2
38	3	3
39	3	3
40	2	2,25
41	3	2,75
42	5	4,75
43	5	4,75
44	4	3,75
45	5	4,75
46	5	4,75
47	4	4,5
48	5	4,5
49	3	2,25
50	3	2,25
Mean	3,56	3,32

4. Discussions, Limitations, and Conclusions

This study aimed to explore the holistic and analytic way of assessing speaking skills in a higher-education setting. Altogether 50 third-year university students undertook an oral examination in the subject English language at a Slovak university. The examination was at the C1 level according to CEFR. Both holistic and analytic ways of scoring were employed.

The results demonstrate that the subjects achieved in the four categories – content and organisation, pronunciation, vocabulary, and grammar, 3.96, 3.54, 3.22, and 2,54 points respectively. Despite the fact that CLT ought to be the primary method of teaching English as a foreign language, it appears that L2 learners experience problems when they have to use B2/C1 level words along with more complex and sophisticated grammatical structures within their utterances. This may be the result of applying the Grammar-translation method (in its various forms) to some degree in Slovak school system of education. The learners might even know the B2/C1 words, but they are not able to use them when they speak. Therefore, following the principles of CLT, and providing EFL learners with more space for practicing speaking might prove useful.

Furthermore, the results reveal that the average score for holistic and analytic ways of scoring was 3.56 and 3.32, respectively. The p value was calculated at 0.001; thus, there exists a statistically significant difference between the holistic and analytic methods of scoring ($p < 0.05$).

This does not necessarily mean that one method of scoring is more reliable than another 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 regarded as useful and appropriate since the two methods seem to complement each other. Moreover, the analytic scoring enabled the subjects to be provided with a rather detailed feedback on their performance in particular categories. Finally, the findings offered useful diagnostic information so that both the EFL higher-education students and their teachers know which areas they should concentrate on more.

There are several limitations to this study. First, there was only one assessor (and one interlocutor) and their subjective perception and interpretation of a candidate's oral performance might have influenced the assessment process. However, it should be noted that the evaluation of spoken proficiency is a highly subjective process, and there are numerous factors which influence the assessor's judgement (Jankowska and Zielińska, 2015). It is thus suggested that future research employs a higher number of assessors in order to provide more statistical power for the evaluation of the relationship between the holistic and analytic ways of scoring. Similarly, a larger sample of subjects can be adopted in future studies too. Furthermore, the description of bands in the analytical scoring scales may have also played its part within the assessment process. Again, a subjective interpretation might have influenced the assessment process. Nonetheless, it should be emphasized that it is a rather difficult task to offer clear-cut and unambiguous definitions for the descriptors (Llach, 2011). It seems reasonable to assume that the level of subjectivity can be decreased by undergoing an appropriate training and by gaining years of experience, so that the assessment can become as accurate, reliable, and objective as possible. Finally, it might be interesting to compare the difference between female and male scores within the assessment of spoken proficiency in future studies.

It can be concluded that combining the analytic and holistic scoring may be regarded as a rather viable option when it comes to the assessment of speaking skills. Both ways have their advantages and drawbacks, and employing these two methods of scoring might possibly result in 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.

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. *CBE— 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. *Unknown Journal* 1633.

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., and 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

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.

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.

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. Waniek-Klimczak (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.

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.

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.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

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

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, & S. Stoynoff (Eds.) *The Cambridge Guide to Second Language Assessment*, 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.

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.

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 AP English Classes. *Journal of Adolescent & Adult Literacy, 60* (1), 13-23. doi: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.

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 (TAST) for Operational Use. *Language Testing, 24* (2), 251-286. https://doi.org/10.1177/0265532207076365.