International Journal of Instruction *January 2019 ● Vol.12, No.1*
e-ISSN: 1308-1470 ● www.e-iji.net *p-ISSN: 1694-609X*
pp.1059-1076

Received: 02/07/2018
Revision: 07/10/2018
Accepted: 12/10/2018
OnlineFirst: 10/12/2018

Malaysian Speaking Proficiency Assessment Effectiveness for Undergraduates Suffering from Minimal Descriptors

Karwan Mustafa Saeed
PhD, Faculty of Education, Koya University, Danielle Mitterrand Boulevard, Koya KOY45, Kurdistan Region, Iraq, *karwanbogdy@yahoo.com*

Shaik Abdul Malik Mohamad Ismail
Assoc. Prof., corresponding author, School of Educational Studies, Universiti Sains Malaysia, Malaysia, *samohame@gmail.com*

Lin Siew Eng
Asst. Prof., Faculty of Social Sciences & Liberal Arts, UCSI University, Malaysia, *linse@ucsiuniversity.edu.my*

This study was primarily aimed at developing an English-speaking proficiency test and analytic rubrics designed to measure speaking proficiency of Malaysian undergraduates. On the basis of Littlewood's Methodological Framework and Long's Interaction Hypothesis, the researchers derived three speaking tasks from four sources: (a) syllabus of the English language courses at the relevant university, (b) Kathleen Bardovi-Harlig's operationalizing conversation speech acts, (c) IELTS part B speaking test, and (d) task B speaking section of Malaysian University English Test (MUET). A total of 96 undergraduates with four levels of the language proficiency (i.e., low performers, intermediate performers, upper-intermediate performers, and high performers) from a public university in Malaysia voluntarily participated in the study. While two TESOL experts were invited to validate the content of the tasks and the rubrics, two raters rated students' test scores. Construct validity was established through a known-group validity (construct validity) for a known-group comparison of the task performance at the three difficulty levels namely, elementary, intermediate and advanced. The test scores, having good internal consistency (a= .89) and inter-rater reliability (ICC= .84), yielded speaking proficiency descriptors. This result showed that the test is reliable and valid to diagnose speaking proficiency of Malaysian undergraduates in pursuit of improvement.

Keywords: speaking assessment, analytic rubrics, validation, reliability, test development

Citation: Saeed, K. M., Ismail, S. A. M. M., & Eng, L. S. (2019). Malaysian Speaking Proficiency Assessment Effectiveness for Undergraduates Suffering from Minimal Descriptors. *International Journal of Instruction*, 12(1), 1059-1076.

INTRODUCTION

In the 21st century, English is regarded as an individual asset for tackling with very competitive job markets (Jiang, 2003). The ability to speak English fluently is the goal for majority of English learners (Mohammadi & Enayati, 2018). How fluently the learner speaks gives the first impression of speaking proficiency (Nunan, 1991) and about an opportunity for employment. As highlighted by Rao and Abdullah (2007) and Simion (2012), to secure job employment in a competitive environment across countries, specifically Malaysia, students need to communicate in English efficiently. They also concluded that only those who possess a reasonably good command of the English language are prefered in the job market. The lack of speaking proficiency among Malaysian graduates is therefore a cause for worry (Lan, Khaun, & Singh, 2011). Notwithstanding that Malaysian students learn English for years at the primary, secondary and tertiary levels, they leave universities with little fluency in speaking English language (Hiew, 2012). This unsatisfactory result has raised the researchers' concern over how to diagnose undergraduates' speaking proficiency during academic years. This concern further gives rise to the question: How effective is the current test of speaking proficiency?

To assess Malaysian undergraduates' language proficiency, MUET has been conducted by the Malaysian Examinations Council since 2000. MUET is aimed at helping stakeholders to assess the overall language level of candidates required to attain a particular band score out of six bands (Malaysian Examinations Council, 2015). However, MUET only provides general descriptions of bands. For example, band 4 description of MUET indicates that candidates "lack the ability to convey the message accurately" but are at the same time "satisfactorily expressive and fluent . . . with occasional inaccuracies" (Malaysian Examinations Council, 2015, p.10). Other bands also have similar contradictory descriptions. Apparently, the MUET speaking assessment rubrics provide little help for differentiating between proficiency levels and provides minimal descriptors for speaking proficiency of the language learners. The language lecturers have little information to design their instructional materials in accordance with the needs of the language learners.

This drawback is not only peculiar to MUET/the local context, but also to band descriptors of rubrics in international contexts. The Common European Framework of Reference for Languages (CEFR), which is also used in non-European countries (Little, 2007), needs more analytic rubrics. Band descriptors of CEFR speaking assessment rubrics have been criticized for ambiguities and inconsistencies about differentiating between proficiency levels (Alderson, 2007; Galaczi, 2013) and suitability for young learners (Hulstijn, 2010; Little, 2007). Similar critique is applicable to traditional assessments that are based on grades or percentages (i.e., only revealing who among students are better than others), which provides no insight or clue on how to improve language proficiency (Kubiszyn & Borich, 2010).

Speaking is one of the most challenging language skills to assess, mainly because it requires to teach individual learners and to assess speaking performance of each individual (Bachman & Palmer, 1996; Luoma, 2004). This challenge could be a reason

that speaking assessment has not been given due attention in universities across countries, instead, the immense focus has been on grammar, and vocabulary (Egan, 1999). Such traditional assessments fall short of gauging a specific aspect of language like speaking (Oosterhof, 2001).

Thus, like the traditional assessment, MUET carries at least two major drawbacks for speaking proficiency during the language teaching and learning process at public universities in Malaysia. First, it provides scarce help to identify strengths and weaknesses of the language learners. Second, its rubric lacks specific descriptors for an accurate interpretation and implication of raw scores. Dealing with these drawbacks requires a diagnostic approach that allows identifying strengths and weaknesses in speaking proficiency. In Malaysia, however, such a diagnostic approach, especially in assessment of speaking proficiency among undergraduates, has yet to be developed. Lecturers in Malaysian public universities usually use holistic scoring to assess speaking proficiency in the classroom. Holistic methods are unable to pinpoint the specific weaknesses of students. Although this method provides the language lecturers with test scores indicating task performance, it remains insufficient for improvement of the language teaching and learning.

As a recent review of educational measurements (Masters, 2015) highlighted, to enable the language instructors to modify teaching materials and strategies for improvement, the diagnostic approach, defined as "formative assessment or assessment for learning", is the most practical approach among others. Therefore, in today's teaching and learning pedagogy, 'assessment of learning' has been replaced by 'assessment for learning' (Khodabakhshzadeh, Kafi & Hosseinnia, 2018). This shift from assessment of learning to assessment for learning (i.e., an alternative diagnostic approach to the traditional assessment) allows developing rubrics with more specific descriptors for serving the next stage of learning (Masters, 2013). In other words, assessment of speaking proficiency should be for the improvement of teaching and learning not merely of the task performance (Alberola Colomar, 2014). To this end, the researchers of the present study developed and tested a new speaking proficiency test (ranging from elementary to advanced levels of tasks) and proposed a new speaking assessment rubric with a new set of descriptors for each band, to diagnose undergraduates' speaking proficiency.

METHOD

Test Design

This quantitative study provides preliminary evidence for designing and validating a prototype speaking test and its assessment rubrics to diagnose undergraduates' speaking proficiency. The test development was based on Littlewood's Communicative Methodological Framework (1981) and Long's Interaction Hypothesis (1981). Applying the recommendation by Littlewood (1981), the test was constructed from pre-communicative activities to free communicative activities at elementary, intermediate, and advanced levels of task difficulty. For the elementary level, ten written-for-oral tasks were used, similar to a number of studies (e.g., Cohen & Shively, 2007; Eslami & Liu, 2013), requiring language learners to write what they would say in conversation.

The elementary tasks were based on Kathleen Bardovi-Harlig's (2015) operationalizing conversation speech acts. The intermediate level required verbal answers to five questions about a written stimulus (i.e., job application interview), which was inspired by part B of IELTS speaking test. To ensure that participants were familiar with the stimulus, it was adapted from the syllabus of the English language courses for undergraduates at the Language Centre of the target University. As for the advanced level, a group discussion was formed according to the Interaction Hypothesis (Long, 1981) and related literature (Ellis, 1999). The discussion topic was adopted from task B of MUET speaking section. Figure 1 illustrates the steps of developing and validating the speaking test.

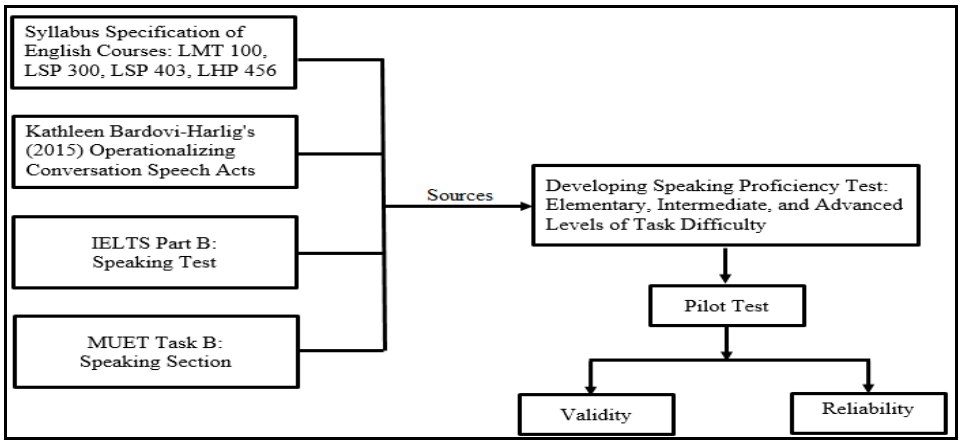


Figure 1
Steps of developing the prototype speaking test

Rubrics for Assessment

As a grading tool, a rubric is an explicit set of criteria, used for assessing a specific task-performance. Rubrics help examiners deal with issues related to assessment, such as reducing grading time, grading more objectively (reducing subjectivity), providing students with timely feedback, and identifying deficiencies in students' learning or performance (Stevens & Levi, 2005). Parts of a grading rubric are criteria, scores, bands (performance levels), and descriptors.

A reliable assessment of speaking proficiency has been a concern to better inform pedagogy and facilitate the learning of speaking (Bachman & Palmer, 1996). To assess speaking task performance, the present researchers designed a speaking rubric based on Canale and Swain's (1980) communicative approach to second language teaching and testing (See Appendix B). The rubric developers were two experts (those who validated the speaking test) and a PhD candidate in the field (language instruction and assessment) at a public university in Malaysia. The experts checked and validated the rubric content.

The communicative approach comprises four areas of competence (i.e., grammatical, sociolinguistic, strategic, and discourse), which can be used as a guide to develop

assessment criteria for a specific purpose, but not necessarily to assess all the areas at a time by a given task, or given equal importance (Chambers & Richards, 1992). Malaysia's Ministry of Higher Education has prioritized grammatical and sociolinguistic competence over the other competencies, mainly because of their importance in pursuing the career, either academic or other professions after graduation. Accomplishing a given task requires not only grammatical competence but also appropriate use of the language within the social contexts (News and business analysis for Professionals in International Education, 2018, para. 4). Hence, this study primarily focused on the assessment of grammatical and sociolinguistic competence.

Criteria and Scoring

To score speaking tests, two basic methods are usually used, namely, holistic (impressionistic) and analytic (Taylor, 2011). The former expresses "an overall impression of an examinees' ability in one score" (Luoma, 2004, p. 61), whereas the latter consists of a number of criteria, each criterion has descriptors for performance levels (Luoma, 2004). The analytical scoring method allows to give a single score for different criteria of speaking proficiency (e.g., vocabulary, communicative activity, pronunciation, and fluency), providing specific assessment for language teaching and learning improvement. It enhances reliability of the assessment (Srikaew, Tangdhanakanond & Kanjanawasee, 2015). Although the analytical rating score is more time-consuming, it is widely used to assess learner performance on authentic language speaking assessments (Fulcher, 2003). Therefore, as Srikaew et al. (2015) suggested, the analytical rating score is essential for a highly reliable and fairer assessment.

The assessment criteria covered appropriateness of speech, communicative ability, managing discussion, fluency, pronunciation, grammar, and vocabulary, and the criteria were matched with existing scoring rubrics to design a scoring rubric template. These criteria for assessment of speaking proficiency are not equally weighted (marked) in speaking assessment (i.e., some criteria receive higher award than others). This analytical assessment is referred to as weighting system (Underhill, 1987), a procedure in which marks are given out of the same total initially (i.e., out of the same maximum mark instead of marking one criterion out of ten, another out of twenty, at the same time) and then multiplied by different factors to obtain a weighted score (e.g., appropriateness is marked out of 5 then multiplied by 3). Marks are awarded according to the degrees of correctness, called partial-credit scoring (Bachman & Palmer, 1996) by giving numbers of 0 (no evidence of knowledge), 1 (evidence of very limited knowledge), 2 (evidence of limited knowledge), 3 (evidence of moderate knowledge), 4 (evidence of extensive knowledge), and 5 (complete evidence of knowledge). The marking of the students' performance was based on a weighting system and partial-credit scoring as shown in Table 1.

Table 1
Criteria for the speaking assessment

Levels	Criteria	Marks
Elementary Level	Appropriateness	marked out of 5 then multiplied by 3
	Grammar	marked out of 5 then multiplied by 2
Intermediate Level	Appropriateness	marked out of 5 then multiplied by 3
	Communicative ability	marked out of 5 then multiplied by 3
	Fluency	marked out of 5
	Pronunciation	marked out of 5
	Grammar	marked out of 5
	Vocabulary	marked out of 5
Advanced Level	Communicative ability	marked out of 5
	Fluency	marked out of 5
	Pronunciation	marked out of 5
	Grammar	marked out of 5
	Vocabulary	marked out of 5

Bands/Levels of Speaking Performance and Descriptors

An overall raw score is used to identify whether the language learner has higher or lower performance compared with other test takers, but it falls short of identifying the learner's performance band, provides no descriptors of the proficiency level (Kubiszyn & Borich, 2000). To identify the performance bands descriptors, criterion-referenced cut-off scores are helpful. Criterion-referenced tests are measures that allow to ascertain task performance of a test taker with respect to a set of criteria rather than a comparison with other testees (Popham & Husek, 1969). Further, a criterion-referenced test as an authentic and productive test, is the answer to the need of language teaching and testing communicatively (Wullur, 2011). Educational tests for instructional decision making are often criterion referenced. For example, criterion referencing is used to identify course content that a student has and has not mastered, so that deficiencies can be addressed before moving forward (Albano, 2016). A set of criteria can be established by a panel of experts who can determine categorizing students into performance bands, that is, students' task performance can be assessed according to the description of each performance band (Albano, 2016).

Luoma (2004) posits that choosing the number of bands is an essential concern to distinguish between the bands. As shown in Table 2, to categorize the participants' raw scores according to corresponding descriptors, the four raters (two for content validation and two for test scores) suggested four bands (as in Weir, 1993, as cited in Weir, 2005). Selecting four levels of differentiation seems more practicable for reliability, considering the time constrains whereas fewer than four bands would not be feasible for differentiation between learners and more than six bands would become difficult to differentiate consistently (Luoma, 2004). The raters believed that the categories and the performance bands would provide appropriate feedback for teaching and learning.

Table 2
Levels of performance bands

Bands	Proficiency levels	Scores
Band 1	Novice Learners	Between (0-25)
Band 2	Intermediate Learners	Between (26-50)
Band 3	Advanced Learners	Between (51-75)
Band 4	Superior Learners	Between (76-100)

Descriptions of the performance levels for each criterion of speaking proficiency describe how well the learners perform and what performance at each specific level looks like. The same descriptors have been used for the different criteria of speaking proficiency within the rubrics. To distinguish between the four proficiency levels, Novice Learners hardly speak as in band 1, Intermediate Learners speak with difficulty as in band 2, Advanced Learners speak satisfactorily as in band 3, and Superior Learners speak very well as in band 4.

Participants

A total of 96 undergraduate students (aged 19-23 years, males=24 and females=72) voluntarily participated in the study. They were all in their first through fourth year of their studies in different academic disciplines at public university in Malaysia. They were enrolled in different levels of English language courses at the Language Centre of the university for improvement of the language proficiency. The participants have to attempt the elementary and intermediate levels individually but for the advanced level, they will have to carry out a group discussion.

Data Collection Procedure

Data were collected upon receiving the administrative approval and participants' consent. The data collection from the volunteer participants took four days. Each day 24 participants gathered in a secured room and completed the elementary task. However, to avoid possible influence of exposure among participants on their performance of the same intermediate and advanced tasks, separate rooms were allocated for those who performed and those who were yet to perform the tasks. While they were individually called to complete the intermediate task, they were randomly grouped in four to actively participate in the group discussion/the advanced task. The intermediate speaking task was audio-recorded, whereas the advanced speaking task was video-recorded (Underhill, 1987).

Test Administration Time

In general, there is no standard time for speaking tests. For example, IELTS speaking test takes up to 14 minutes whereas the speaking section of the TOEIC takes 20 minutes, and MUET speaking section 12 minutes to complete. This difference in test administration time depends on test specifications. In this study, different time was allocated to different levels of speaking task preparation and completion. There was no preparation time for the elementary level, but one minute for the intermediate and two minutes for the advanced level. The maximum time for the test completion was 10

minutes for the elementary, 3 minutes for the intermediate, and 10 minutes for the advanced level. As prior research (Hirai & Koizumi, 2009) also revealed, unlike upper-intermediate and higher performing, those lower-performing participants tended to produce less extended talk at the intermediate and advanced tasks and took extra time for completing the elementary task as shown in Table 3.

Table 3
Test time administration

English proficiency level performers	Time taken		
	Elementary Level	Intermediate Level	Advanced Level
Low Performers	12 minutes	2 minutes	7 minutes
Intermediate Performers	12 minutes	2 minutes	7 minutes
Upper-intermediate Performers	10 minutes	3 minutes	9 minutes
High Performers	10 minutes	3 minutes	9 minutes

FINDINGS AND DISCUSSION

This research set out to develop a valid and reliable speaking proficiency test and a new specific speaking proficiency assessment rubric to assess the undergraduates' speaking proficiency performance. Therefore, reliability and validity of a test is central to its consistency and accuracy for either subjective or objective assessment (Krzanowski & Woods, 1984). Especially for a speaking test, subjective ratings or assessment by raters should be reliable (Sawaki, 2007). The proposed speaking test was assessed for its internal consistency, inter-rater correlation coefficient, and validity.

One method for estimating reliability of a test is the Cronbach's Alpha analysis, which requires no dichotomous score (Gay & Airasian, 2003). Using IBM-SPSS-Version-23, the Cronbach's Alpha was generated for estimating internal concistency of the test. The alpha value of .89 indicated a strong reliability.

The validity of scoring rubric can be established based on its reliable application, and this can be examined through scoring consistency (Luoma, 2004), and scoring consistency (rater reliability) is estimated from the perspective of inter-rater reliability (Weir, 2005). To this end, two raters were introduced to the rubric in which they had to learn how to score the speaking tasks. They scored the participants' speaking performance independently. Their scores were estimated through intraclass correlation coefficient (ICC), which is a measure of inter-rater reliability, with the coefficient value of .70 as acceptable, above .80 as good, and above .90 as excellent (Linn & Miller, 2005). ICC was estimated based on absolute-agreement, 2-way mixed-effects model, 95% confident intervals. There was a high degree of reliability of ratings by the two raters (Table 4). The average measure ICC was .844 with a 95% confidence interval from .766 to .896 ($F = 6.371$, $p<.001$).

Table 4
ICC estimates

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Average Measures	.844	.766	.896	6.371	95	95	.000

Validity of a test depends on whether it measures what it is supposed to measure (Kubiszyn & Borich, 2000), such as group differences (DeVon et al., 2007). Such a test must discriminate across groups that are theoretically known to differ, referred to as "known-group" validity (Hattie & Cooksey, 1984), which is a form of construct validity (DeVon et al., 2007; Portney & Watkins, 1993). The test validity is established by a statistical comparison of mean scores across groups (MacKenzie et al., 2011). As Cronbach and Meehl (1955) stated: "If our understanding of a construct leads us to expect two groups to differ on the test [scale], this expectation maybe tested directly" (p.287).

In this study, high-performing students were theoretically expected to perform higher on the speaking test than of those lower-performing students. Therefore, to "empirically" distinguish between high and low performers, known-group validity of the test was assessed across four groups (four levels of the language proficiency) at the three difficulty levels of the speaking test. The mean score of the four groups at each level was calculated (Table 5). The lower-performers performed lower in the three tasks compared with those of higher-performers. Further, the same group performance showed a gradual decrease across the three levels of task difficulty, indicating a difference in speaking performance within and between the groups. Hence, construct validity of the test is promising.

Table 5
Mean scores of respondents' speaking proficiency

Participants	Mean		
	Elementary task (25)	Intermediate task (50)	Advanced task (25)
Low Performers	12.08	18.25	8.25
Intermediate Performers	15.25	23.04	10.41
Upper-intermediate Performers	18	28.79	12.83
High Performers	20.58	33.50	15.45

To secure a job in this era anywhere around the world, including Malaysia, university graduates are expected to possess a good command of English, especially speaking proficiency. Hence, the lack of speaking proficiency among Malaysian graduates needs to be addressed (Lan, et al., 2011). Despite all the efforts taken by the Ministry of Higher Education and learning English for several years in schools, university graduates have yet to master English speaking proficiency (Hiew, 2012). A possible cause for this undesirable result is how speaking assessment has been conducted during university academic years. As Bachman and Palmer (1996) and Luoma (2004) highlighted,

assessing speaking proficiency of language learners is a difficult task because there are several factors that influence our understanding of how well an individual can speak a language, and because test scores are expected to be accurate and appropriate for the intended purpose. Luoma (2004) further explains that "from a testing perspective, speaking is special because of its interactive nature, and it is often tested in live interaction, where the test discourse is not entirely predictable, just as no two conversations are ever exactly the same even if they are about the same topic and the speakers have the same roles and aims" (p.170). This challenge could be a reason that speaking assessment has not been paid close attention in universities, instead, massive focus has been on grammar, vocabulary, and written tests. Such traditional assessments provide no help to identify strengths and weaknesses of the students. Further, traditional rubrics (e.g., MUET speaking rubrics) lack specific descriptions of their bands. Addressing these drawbacks is an important issue aiming at improving students' speaking proficiency.

The result of this study is a proposed test and speaking proficiency assessment rubrics. The test is categorized in three levels, namely; elementary, intermediate and advanced. The proposed test and rubric are designed to provide language instructors with a valid and reliable criterion-based assessment to be used in universities to help identify the strengths and weaknesses of the language learners, thereby helping them to improve their speaking proficiency. Research on language teaching and learning has reached a consensus that language should be taught and assessed on the basis of communicative activities for the improvement of speaking proficiency (Bachman & Palmer, 2010; Canale, 1983; Canale & Swain, 1980). The current study, therefore, attempted to enhance upon the existing measures of speaking proficiency assessment in order to make speaking assessment more accurate.

Findings of the construct validity indicated that the developed test was able to distinguish between higher and lower performing students. This is an interesting fact that the test was able to distinguish between students of different language proficiency levels. This finding showed parallelism with what Fulcher (2003) found, claiming that task difficulty is related to construct validity of a language test. Likewise, findings of the Cronbach's Alpha and inter-rater reliability revealed that the proposed test is reliable to assess undergraduates' speaking proficiency. As Hughes (2003, p.42) notes, in the case of criterion-referenced assessment, seventy percent agreement "is an accepted estimate of decision consistency".

The scoring rubrics offered great potentials to yield beneficial backwash for students and teachers. From the students' perspective, the scoring rubric serves as a source of pre-assessment preparation reference and post-assessment analytic feedback, the latter of which can provide students specific information on their underachievement areas of their speaking proficiency. From the teachers' perspective, the scoring rubrics should provide necessary information on instructional objectives which might not have been obtained adequately, thereby using it to improve instructional materials. Therefore, this research has helped to raise more awareness of the aspect of validity in assessment.

IMPLICATIONS

This research holds significant implications for the teaching and assessment of speaking proficiency. First, language lecturers and instructors can use the speaking rubric in speaking courses to assign expectations in the beginning of their speaking instruction to provide feedback on students' improvement. This suggestion appears in line with previous literature, indicating that providing learners with rubrics contributes to their language learning, because rubrics identify areas for improvement in instruction (Fleming, 2001; Song, 2006). Although language lecturers may see rubrics solely as tools for grading and assessment, students report that rubrics help them in improvement (Reddy & Andrade, 2010). Second, the ESL lecturers can tailor the instructional materials needed based on the students' weaknesses in specific areas of speaking in universities because the speaking rubric and descriptors serve as an effective analytical instrument to assess the effectiveness of their instructional strategies and materials for what their students have or have not mastered. A further implication of this research is its contribution to the continuous explication on developing assessment rubric to increase grading reliability. The method adopted in this research provided a probable model for devising of new assessments, which can serve as a model for future scale developers.

CONCLUSION AND FUTURE RESEARCH

Speaking in English fluently and efficiently is a primary objective to educational establishments in Malaysian higher education. This study has argued the underachievement of this objective because Malaysian university graduates are found to be deficient in the expected speaking proficiency level of performance. The review of literature has asserted that assessment plays a central role in student improvement, and as such imprecise or inadequate speaking assessment leads to poor performance in speaking proficiency. Despite some limitations, this study has attempted to provide a valid and reliable speaking proficiency test as well as speaking assessment rubrics to diagnose students' speaking performance and identify the areas where the students are lacking. The study concludes that a criterion-referenced performance test is the answer to the need of student assessment for betterment.

The study is a small scale in nature. Therefore, future studies can address this issue by considering students from different universities. Furthermore, establishing content and construct validity may not be sufficient in itself for a comprehensive validity argument, but it is an essential first step for operational high-stakes tests. Applied studies, therefore, are required to guide the profession in operationalizing validity in all its manifestations. This lies in further modifications and refinement. The construct, including the assessment criteria and the performance bands requires more rigorous think-allowed protocol and more detailed data analysis if greater objectivity, validity and reliability are to be obtained. Suggestions for future research can also include a closer focus on learning and research on speaking assessment rubric use in diverse higher educational contexts. similar measure for further speaking assessment studies. It

is hoped that this study will encourage prospective studies on speaking assessment to advance further with the methodology and make their efforts available.

REFERENCES

Albano, A. D. (2016). EDPS 870: Introduction to educational and psychological measurement—A peer review of teaching project benchmark portfolio. *UNL Faculty Course Portfolios. 29.* Retrieved 03 March, 2018 from http://digitalcommons.unl.edu/prtunl/29

Alberola Colomar, M. P. (2014). A classroom-based assessment method to test speaking skills in English for specific purposes. *Language Learning in Higher Education, 4*(1), 9-26.

Alderson, J. C. (2007). The CEFR and the need for more research. *Modern Language Journal, 91*, 659–663.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford: Oxford University Press.

Barvdovi-Halrig, K. (2015). Operationalizing conversation in studies of instructional effect in L2 pragmatics. *System, 48*(1), 21-34.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 2-27). New York: Longman Press.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics, 1*(1), 1-47.

Chambers, F., & Richards, B. (1992). Criteria for oral assessment. *The Language Learning Journal, 6*(1), 5-9.

Chapelle, C.A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing, 20*(4), 409–439.

Cohen, A. D., & Shively, R. L. (2007). Acquisition of requests and apologies in Spanish and French: Impact of study abroad and strategy-building intervention. *The Modern Language Journal, 91*(2), 189-212.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., ... & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing scholarship, 39*(2), 155-164.

Egan, K.B. (1999). Speaking: A critical skill and a challenge. *Calico Journal, 16*(3),

277-293.

Ellis, R. (1999). *Learninga second language through interaction.* Amsterdam/Philadelphia: John Benjaminds Publisihing.

Eslami, Z., & Liu, C. N. (2013). Learning pragmatics through computer-mediated communication in Taiwan. *International Journal of Society, Culture & Language, 1*(1), 52-73.

Fleming, V. M. (2001). Helping students learn to learn by using a checklist, modified rubrics, and e-mail. *Journal on Excellence in College Teaching, 12*, 5-22.

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

Galaczi, E. (2013). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics, 35*(5), 553-574.

Gay, L. R., & Airasian, P. (2003). *Educational research: Competencies foranalysis and application* (7th ed.). Upper Saddle River, New Jersey: Merril Prentice-Hall.

Hattie, J., & Cooksey, R. W. (1984). Procedures for assessing the validities of tests using the" known-groups" method. *Applied Psychological Measurement, 8*(3), 295-305.

Hiew, W. (2012). English language teaching and learning issues in Malaysia: Learners' perceptions via Facebook dialogue journal. *Journal of Arts, Science and Commerce, 3*(1), 11-19.

Hirai, A., & Koizumi, R. (2009). Development of a practical speaking test with a positive impact on learning using a story retelling technique. *Language Assessment Quarterly*, *6*(2), 151-167.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Hulstijn, J. H. (2010). Linking L2 proficiency to L2 acquisition: Opportunities and challenges of profiling research. In I. Bartning, M. Martin & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp.233–238). Amsterdam: European Second Language Association.

Jiang, Y. (2003). English as a Chinese language. *English Today, 19*(2), 3–8.

Khodabakhshzadeh, H., Kafi, Z., & Hosseinnia, M. (2018). Investigating EFL Teachers' Conceptions and Literacy of Formative Assessment: Constructing and Validating an Inventory. *International Journal of Instruction, 11*(1), 139-152.

Krzanowski, W. J., & Woods, A. J. (1984). Statistical aspects of reliability in language testing. *Language Testing, 1*(1), 1-20.

Kubiszyn, T., & Borich, G. (2000). *Educational testing and measurement: Classroom application and management* (6th ed.). New York, USA: John Wiley & Sons, Inc.

Kubiszyn, T., & Borich, G. (2010). *Educational testing and measurement: classroom application and practice* (9th ed.). New Jersey, USA: John Wiley and Sons, INC.

Lan, C. O. T., Khaun, A. L. C., & Singh, P. K. S. (2011). Employer expectations of language at the workplace. *Malaysian Journal of ELT Research, 7*(2), 82-103.

Linn, R. L., & Miller M. D. (2005). *Measurement and Assessment in Teaching* (9th ed.). Upper Saddle River, New Jersey: Merrill Prentice Hall.

Little, D. (2007). The Common European framework of reference for languages: Perspectives on the making of supranational language education policy. *Modern Language Journal, 91*, 645–655.

Littlewood, W. (1981). *Communicative language teaching: An introduction*: Cambridge University Press.

Long, M. (1981). Input, interaction, and second-language acquisition. *Annals of the New York Academy of Sciences, 379*(1), 259-278.

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

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly, 35*(2), 293-334.

Malaysian Examinations Council (2015). *Malaysian University English Test (MUET): Regulations, test specifications, test format and sample questions*. Selangor, Malaysia. Retrieved 15 March, 2018 from http://www.mpm.edu.my/download_MUET/MUET_Test_Specification_2015VersiPortal.pdf

Masters, G. N. (2013). *Towards a growth mindset in assessment.* Melbourne, Victoria: ACER.

Masters, G. N. (2015). Rethinking formative and summative assessment. *Teacher Magazine*. Retrieved 10 April, 2018 from https://www.teachermagazine.com.au/geoff-masters/article/rethinking-formative-and-summative-assessment

Mohammadi, M., & Enayati, B. (2018). The Effects of Lexical Chunks Teaching on EFL Intermediate Learners' Speaking Fluency. *International Journal of Instruction, 11(*3), 179-192.

News and Business Analysis for Professionals in International Education (2018, January 4). *English learning overhauled in Malaysia* [News Release]. Retrieved 10 April, 2018 from https://thepienews.com/news/mohe-overhauls-english-learning-malaysian-students/

Nunan, D. (1991). *Language teaching methodology.* A textbook for teachers. London: Prentice Hall International.

Oosterhof, A. (2001). *Classroom applications of educational measurement* (3rd ed.). Upper Saddle River, New Jersey, USA: Merril Prentice-Hall.

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement*, *6*(1), 1-9.

Portney, L. G., & Watkins, M. P. (1993). Foundations of clinical research: Applications to practice. Norwalk: Appleton & Lange.

Rao, R., & Abdullah, S. (2007). The role of English language in the tourism industry. *Universiti Utara Malaysia*. Retrieved 21 September, 2018 from http://repo.uum.edu.my/3221/1/Ravi1.pdf

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & evaluation in higher education*, *35*(4), 435-448.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*(3), 355 – 390.

Simion, O., M. (2012). The importance of teaching English in the field tourism in universities. *An Economic Series Journal, 2*(2), 126-214.

Song, K. H. (2006). A conceptual model of assessing teaching performance and intellectual development of teacher candidates: A pilot study in the US. *Teaching in Higher Education*, *11*(2), 175-190.

Srikaew, D., Tangdhanakanond, K., & Kanjanawasee, S. (2015). Development of an English speaking skill assessment model for grade 6 students by using portfolio. *Procedia-Social and Behavioral Sciences*, *191*, 764-768.

Stevens, D. D., & Levi, A. J. (2005). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning.* Sterling, VA: Stylus.

Taylor, L. (2011). *Examining speaking: Research and practice in assessing second language speaking.* Cambridge: UCLES/Cambridge University Press.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*: Cambridge, UK: Cambridge University Press.

Weir, C.J. (2005). *Language testing and validation: An evidence-based approach*. New York, New York: Palgrave MacMillan.

Wullur, B. G. (2011). Developing an English performance test for incoming Indonesian students. *Indonesian Journal of Applied Linguistics, 1*(1), 58-72

Appendix A: TEST TASKS

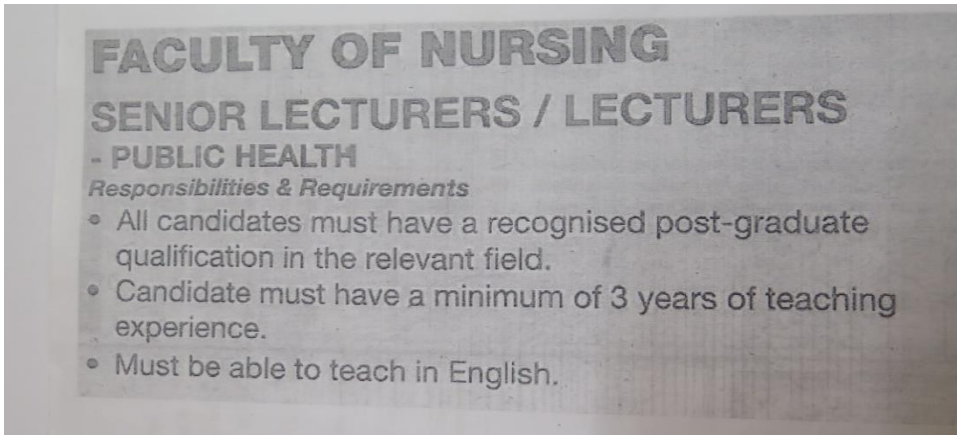
Elementary level

There are ten situations described below. Please read the description of each situation and write down what you would say in that situation.

No	Item
1	You are at your father's office. One of his friends comes over and your father introduces his friend to you. What would you say to your father's friend?
2	You are a student. You forgot to do the assignment for your English course. Your teacher who m you have known for a while now asks for your assignment. You apologize to your teacher. What do you say to him/her?
3	You have a difficult exam tomorrow. You don't understand some of the topics included in the exam. You want to ask one of your friends to help. What do you say to him/her?
4	It is raining hard and you are walking to school. A friend stops his car to offer you a ride. What would you say to him?
5	In a group discussion, your class is discussing spending time on Facebook. One of your classmates believes that nowadays people spend much time on Facebook, and you have the same opinion as your classmate. What would you say to him/her?
6	You went to see a movie at the cinema at the Queensbay Mall, and you loved the movie so much. You believe it was an awesome movie, but your friend, Ali says: The movie was so boring. What would you say to Ali?
7	You are trying to apply to do a master's degree in management in the USA. You are required to provide a recommendation letter from one of your professors. What would you say to your professor to write you a letter?
8	You go to your school library with several books in your hands. Suddenly, you see a librarian. How do you ask him to help you to open the library door for you?
9	You need to talk to your lecturer. You go to his office to know if he has time to talk to you. His office door is open. How do you ask him if he has time to talk to you?
10	You and your friends have been invited by a new friend for dinner. You want to accept your friend's invitation. What would you say to him?

Intermediate level

Read the following flyer and then I will ask you the following five questions.



Intermediate level questions

No	Item
1	How did you know about this teaching job vacancy?
2	Can you tell me about your qualifications for this job?
3	What is your teaching experience in the relevant field?
4	Can you explain how qualified you are for this job?
5	What are your salary expectations? What if we can't fulfil your salary expectations?

Advanced level

Read the following scenario and discuss the question among you. Each one plays a role in the discussion.

Scenario
It has been said that young people in Malaysia today are considered lucky. **Which of the following has helped young Malaysians today the most?** Candidate A: They grew in a time of peace and prosperity. Candidate B: They have easy access to more information. Candidate C: The government has provided better facilities for sports and recreation. Candidate D: The education system has offered them more opportunities.

Appendix B: SPEAKING RUBRICS

Difficulty levels	Assessment criteria	Band 1 (0-29) raw score	Band 2 (30-53) raw score	Band 3 (54-77) raw score	Band 4 (78-100) raw score
Elementary	Appropriateness	• He/she can **hardly** answer in given context appropriately for the intended purpose. • He/she understands questions but can **hardly** perform in good command of form and function.	• He/she **has difficulty** in answering appropriately in given context for the intended purpose. • He/she understands questions but **has difficulty** in good command of form and function.	• He/she answers in given context for the intended purpose **satisfactorily**. • He/she understands questions and has **satisfactory** command of form and function.	• He/she answers **very well** and appropriately in given context for the intended purpose. • He/she understands questions and **very well** command of form and function.
	Grammar	• He/she **hardly** uses accurate and correct grammar**.**	• He/she **has difficulty** in using accurate and correct grammar.	• He/she uses accurate and correct grammar **satisfactorily**.	• He/she uses accurate and correct grammar **very well**.
Intermediate	Appropriateness	• He/she understands questions, but **hardly** speaks appropriately in given context for the intended purpose. • He/she can **hardly** answer interview questions.	• He/she understands questions but **has difficulty** in speaking appropriately in given context for the intended purpose. • He/she **has difficulty** in answering interview questions.	• He/she understands questions and speaks appropriately in given context for the intended purpose **satisfactorily.** • He/she answers interview questions **satisfactorily**.	• He/she understands questions and speaks appropriately in given context for the intended purpose **very well**. • He/she answers interview questions **very well.**
	Communicative ability	• He/she is **hardly** able to answer questions meaningfully. • He/she is **hardly** able to demonstrate well in conveying his/her message.	• He/she **has difficulty** to answer questions meaningfully. • He/she **has difficulty** to demonstrate well in conveying his/her message.	• He/she is **satisfactorily** able to answer questions meaningfully. • He/she is able to demonstrate well in conveying his/her message **satisfactorily**.	• He/she is able to answer questions **very well**. • He/she is able to demonstrate **very well** in conveying his/her message.
	Fluency	• He/she can **hardly** speak fluently and smoothly. • He/she can **hardly** speak without any pausing for too long and connecting his/her ideas.	• He/she **has difficulty** in speaking fluently and smoothly. • He/she **has difficulty** in speaking without any pausing for too long.	• He/she speaks fluently and smoothly **satisfactorily**. • He/she speaks without any pausing for too long **satisfactorily**.	• He/she speaks fluently and smoothly **very well**. • He/she speaks without any pausing for too long **very well**.

	Pronunciation	• He/she **hardly** pronounces the individual words correctly. • He/she is **hardly** able to express stress and intonation correctly.	• He/she **has difficulty** to pronounce the individual words correctly. • He/she **has difficulty** to express stress and intonation correctly.	• He/she pronounces the individual words **satisfactorily**. • He/she is **satisfactorily** able to express stress and intonation correctly.	• He/she pronounces the individual words **very well**. • He/she is able to express stress and intonation **very well**.
	Grammar	• He/she **hardly** uses a range of accurate and correct grammar.	• He/she **has difficulty** to use a range of accurate and correct grammar.	• He/she uses a range of accurate and correct grammar **satisfactorily**.	• He/she uses a range of accurate and correct grammar **very well**.
	Vocabulary	• He/she **hardly** uses a wide range of vocabulary effectively. • He/she **hardly** uses appropriate vocabulary.	• He/she **has difficulty** to use a wide range of vocabulary effectively. • He/she **has difficulty** to use appropriate vocabulary.	• He/she uses a wide range of vocabulary **satisfactorily**. • He/she uses appropriate vocabulary **satisfactorily**.	• He/she uses a wide range of vocabulary effectively **very well**. • He/she uses appropriate vocabulary **very well**.
Advanced	Communicative ability	• He/she is **hardly** able to communicate effectively with the other candidates. • He/she is **hardly** able to demonstrate good interactive ability in carrying out the discussion and maintain eye contact.	• He/she **has difficulty** to communicate effectively with the other candidates. • He/she **has difficulty** to demonstrate good interactive ability in carrying out the discussion and maintain eye contact.	• He/she is able to communicate **satisfactorily** with the other candidates. • He/she is able to demonstrate interactive ability in carrying out the discussion and maintain eye contact **satisfactorily**.	• He/she is able to communicate effectively with the other candidates **very well**. • He/she is able to demonstrate interactive ability in carrying out the discussion and maintain eye contact **very well**.
	Fluency	• He/she can **hardly** speak fluently and smoothly. • He/she can **hardly** speak without any pausing for too long.	• He/she **has difficulty** in speaking fluently and smoothly. • He/she **has difficulty** in speaking without any pausing for too long.	• He/she speaks fluently and smoothly **satisfactorily**. • He/she speaks without any pausing for too long **satisfactorily**.	• He/she speaks fluently and smoothly **very well**. • He/she speaks without any pausing for too long **very well**.
	Pronunciation	• He/she **hardly** pronounces the individual words correctly. • He/she is **hardly** able to express stress and intonation correctly.	• He/she **has difficulty** to pronounce the individual words correctly. • He/she **has difficulty** to express stress and intonation correctly.	• He/she pronounces the individual words **satisfactorily**. • He/she is **satisfactorily** able to express stress and intonation correctly.	• He/she pronounces the individual words **very well**. • He/she is able to express stress and intonation **very well**.
	Grammar	• He/she **hardly** uses a range of accurate and correct grammar.	• He/she **has difficulty** to use a range of accurate and correct grammar.	• He/she uses a range of accurate and correct grammar **satisfactorily**.	• He/she uses a range of accurate and correct grammar **very well**.
	Vocabulary	• He/she **hardly** uses a wide range of vocabulary effectively. • He/she **hardly** uses appropriate vocabulary.	• He/she **has difficulty** to use a wide range of vocabulary effectively. • He/she **has difficulty** to use appropriate vocabulary.	• He/she uses a wide range of vocabulary **satisfactorily**. • He/she uses appropriate vocabulary **satisfactorily**.	• He/she uses a wide range of vocabulary effectively **very well**. • He/she uses appropriate vocabulary **very well**.