



**LEARNER-DRIVEN ORAL ASSESSMENT CRITERIA FOR ENGLISH
PRESENTATION**

¹Mardiana binti Idris & ^{*2}Abdul Halim bin Abdul Raof

¹ English Language Unit, Kolej Matrikulasi Kejuruteraan Johor, 82000 Pontian, Johor,
Malaysia

²Language Academy, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.

*Corresponding author: m-halim@utm.my

Received: 02 January 2019, Accepted: 02 June 2019

ABSTRACT

Learner-centred assessment has been widely propagated in learner-centred approach. However, learners are rarely given the opportunity to engineer their own assessment. Therefore, this study attempted to gauge (1) the functionality of learner-driven oral assessment criteria scaling structure and (2) the reliability of learner-assessors in applying their own assessment criteria during oral presentation. In this study, 11 participants from an electrical engineering group, which consists of one year programme matriculation students, participated in assessment criteria development. First, participants discussed suitable criteria and scaling structure in small groups. Secondly, each group presented their oral assessment criteria for peer feedback. Thirdly, participants discussed and finalised the oral assessment criteria for the class. Fourthly, to test the learner-driven assessment criteria, three speakers from the group volunteered to present their speech. While presenting, these speakers were assessed by their peers. Participants' ratings and scores were later analysed using the Many-Facet Rasch Measurement (MFRM) software. Findings show that despite the criteria being developed by learners, the scaling structures were functioning usefully with the Rasch Threshold measure indicated more than 1.4 logits between assessment levels and the

learner-assessor reliability was > 0.80. The significance of this study lies in raising awareness for improving learners' oral presentation skills as well as developing learner autonomy.

Keywords: Learner autonomy, learner-centred, oral skills, Rasch measurement.

Cite as: Idris, M. & Abdul Raof, A. H. (2019). Learner-driven oral assessment criteria for English presentation. *Journal of Nusantara Studies*, 4(1), 365-383.
<http://dx.doi.org/10.24200/jonus.vol4iss1pp365-383>

1.0 INTRODUCTION

In recent years, an increasing attention has been directed towards learner-centred approach to curriculum development and instruction in English language teaching (ELT). However, learner-centred assessment approach seems to have been lagging behind in taking its equal stand in the curriculum-instruction-assessment triad of language learning. Nunan (1997) defines learner-centred curriculum as 'a collaborative effort between teachers and learners' (p. 135). Unfortunately, this collaborative effort has yet to be materialized fully as learners may be involved in deciding the content of the curriculum and how it is taught but Little (2005) observed that they are normally 'excluded from the process of evaluating curriculum outcomes, including their own learning achievement' (p. 329). This was also observed by Spiller (2012) who remarked that teachers control the power and choices in assessment which consequently, 'limit the potential for learner development' (p. 2).

In Malaysia, for example, learners have been generally conditioned to depend solely on teachers' assessments rather than their own. As a result, learners as well as teachers may have their own reservations in conducting learner-centred assessment approach in language classrooms. In fact, these reservations could have stemmed from the observation made by Hamidi (2010) that:

The shift from product-oriented approaches to process-oriented approaches to assessment has very soon placed a lot greater demands on learners, teachers, parents, teacher trainers and developers, administrators, curriculum/materials developers, communities, and in short on all those in the state, district, and school levels (p. 5).

Considering that assessment process in learner-centred paradigm has been rarely researched, this study was initiated based on the following objectives:

1. to measure the functionality of learner-driven oral assessment criteria scaling structure, and
2. to gauge the reliability of learner-assessors in applying their own assessment criteria during oral presentation.

In this paper, the literature review will first focus on the development of learner-centred assessment, specifically on theories and pedagogical application. Oral assessment criteria as well as scaling structure and the intricacies of oral skills will be discussed. Then, the study is described and the findings are discussed based on the objectives listed. Implications of the study are also presented and finally, the conclusion offers suggestions for future studies.

2.0 LITERATURE REVIEW

2.1 Learner-Centred Assessment

The theory and pedagogical rational for more learner-centred approaches to teaching has been well-developed decades ago (Reinders, 2010; Aslan & Reigeluth, 2015). In fact, the theory and practice of this approach are generally applied in developing curriculum as well as learning instructions. What seems obvious is the lack of using the same theory and rational for developing learner-centred assessment. This somehow baffles the researcher as assessment should be a part of the learning triangle which merges curriculum, instruction and assessment. According to Keppell and Carless (2006), learning-oriented assessment focuses on learning as its core as well as ‘reconfiguring assessment design so that the learning function is emphasized’ (p. 181). Unfortunately, learner-centred approach to learning seems to concentrate mostly on the theory, curriculum design and instruction and somewhat marginalizes the need for incorporating assessment for effective learning. Shepard (2000) produced a historical overview of the changing paradigm that focuses on aligning curriculum, learning theory and assessment. This is illustrated in Figure 1.

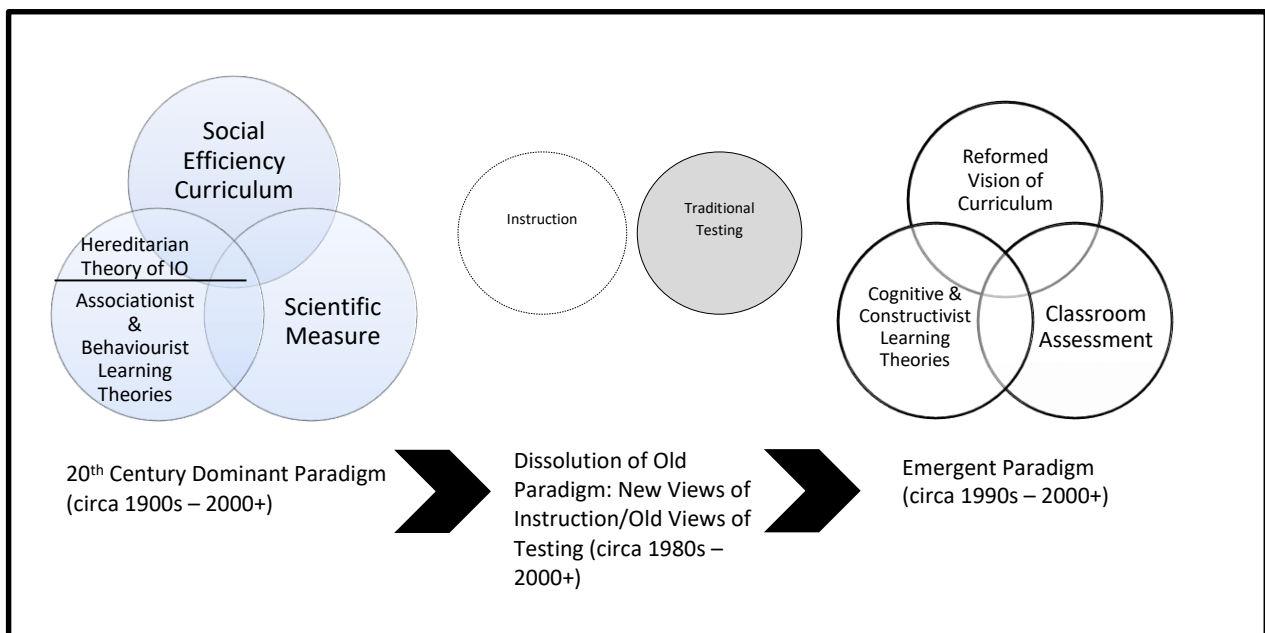


Figure 1: A historical overview illustrating the changing conceptions of curriculum, learning theory and measurement (Shepard, 2000)

Figure 1 illustrates the emergent paradigm of interdependency of curriculum, learning theories and classroom assessment. However, Graue (1993) observed that some language practitioners and curriculum designers view assessment and instruction as conceivably ‘separate in both time and purpose’ (p. 291). This in turn affects the measurement approach to classroom assessment in which standardized tests and ‘teacher-made emulations of those tests’ present a barrier to the implementation of more constructivist approaches to instruction (Shepard, 2000, p. 4). Assessment process in a learner-centred approach seems to be excluded and perhaps ignored due to these reasons:

1. Learners are not able to produce valid assessment criteria due to their lack of knowledge of expertise in the subject being tested.
2. Learners may not be reliable to assess their peers’ performance since they may be influenced by external factors such as emotions.
3. Learners may not be able to apply their own developed oral assessment criteria accurately.
4. Learners may feel that teachers are not doing their jobs.

5. Teachers also require sufficient time in training learners to become assessors and conflict may arise as teachers are under pressure in adhering to the syllabus prescribed.
6. Teachers may feel that learners are taking over their jobs.

Most researchers questioned whether learners are able to produce a functional rating scale especially for oral assessment criteria. In addition, since learners are still learning, the questions of validity and reliability may arise. The researcher is aware of the issue of validity and it is particularly vital in assessment of learning (AoL) and assessment for learning (AfL). However, in assessment as learning (Earl, 2013) or AaL - where the key assessor is the learners themselves - the issue of validity could have been perceived in a different light. Validity refers to how well a test measures what it is purported to measure. Obviously, most learners who have not been exposed to learner-centred assessment approach may find this challenging since the question is: do they really know what needs to be measured? Even if they do know what to measure, are they able to produce relevant or corresponding criteria which measure their oral skills? For instance, are they able to measure accuracy, fluency and complexity of their speaking skills? If the assessment was framed within AoL and AfL, it is understandable for stakeholders of the assessments to raise the issue of validity since the scores awarded may not be valid due to their lack of experience, knowledge and skills in assessing. Since validity is perceived within AaL, perhaps learners' validity issue could be debated from their own understanding of what they need to measure and how this understanding is translated into their own subsequent progress or improvement.

Although engaging learners with assessment criteria fosters meaningful learning (Gikandi, 2011) as prescribed by AaL, this has not been widely practised in language classroom as assessment is still perceived as a product rather than a process and Taras (2008) argued that two pedagogic practices that were conducive to learning were 'discussing and understanding criteria and providing feedback'. Though engagement with assessment criteria differs in degree: ranging from students are 'informed' about it, it is 'discussed' with them, they 'participate' in it to it is 'their responsibility' (Gielen, Dochy, Onghena, Struyven, & Smeets, 2011), positioning and directing learners as active agents in 'assessment decisions' (Boud, Lawson, & Thompson, 2013) can lead to meaningful and sustained learning. A study of learner-oriented assessment practice and oral proficiency by Lim (2007) found that the activity led learners to focus on specific criteria when learning, which consequently enabled better performances. In the study, some learners (6 out of 12) were motivated and found their own

weaknesses (e.g. grammatical mistakes, pronunciation, etc.) and a new way to assess their own language (performance) ability. Another study by Vickerman (2009) that investigated learners' perspectives on formative assessment and its effects to deep approach to learning found that 55% of students agreed or strongly agreed that involvement in the formative peer assessment process enhanced their understanding of assessment criteria. Whilst there was over half of the group who had found the process useful, students who commented to the contrary indicated they found it difficult at times to assess their peers work, and they would have preferred to have more tutor intervention.

2.2 The Intricacies of Oral Assessment Criteria

Assessing learners' oral skills is challenging and letting learners to produce their own oral assessment criteria might have seemed inconceivable. Most studies concur that even for untrained raters, two criteria seem to be easily identified: accuracy and fluency. Accuracy is related to grammar while fluency is related to spontaneous production of the language. However, there are many criteria that need to be looked into while assessing oral skills. Hence, the question is: Are learner-assessors able to identify other criteria apart from accuracy and fluency?

Since assessing learners' oral skills involves observing diverse criteria as well as using the assessment criteria for different purposes (i.e. diagnostic, certification, placement), there are numerous oral assessment criteria which have been developed to serve these purposes all over the world. Some of the oral assessment criteria are easily available such as the following tests: Test of English as a Foreign Language (TOEFL iBT®), International English Language Testing System (IELTS), American Council on the Teaching of Foreign Languages (ACTFL) and Malaysian University English Test (MUET).

Table 1 below shows that each test has a different set of criteria in gauging learners' oral skills and these criteria do not exceed four, even though each criterion may not be similar in weighting and focus. This is consistent with the recommendation that 'four or five categories begin to cause a cognitive load for raters and seven is a psychological upper limit' (Luoma, 2004). Although there is obvious oral proficiency categorisation, one criterion seems to be consistently included in these tests such as accuracy (IELTS and ACTFL) while other tests presumably include this feature within its language criterion (TOEFL and MUET). The stipulated scoring level also varies from four to eleven. In terms of test format, only MUET includes a group discussion of four candidates while the rest test the candidates through

recorded response and interview. Reliability of scores is normally attributed to consistency of judgment between examiners and candidates' scores (Gardner, 2012). From the table, reliability issues may arise in IELTS and ACTFL since these tests only employ one examiner for the interview. Despite repeated claims that the examiner is highly trained, accredited and certified professionally, concerns may surface which pertain to psychological, emotional and environmental states of the examiner at the time of the interview.

Table 1: Assessment criteria of TOEFL iBT®, IELTS, ACTFL and MUET

Test	Orientation	Format	Scoring level	Examiner	Assessment criteria
TOEFL iBT®	Language knowledge	Recorded response	4	3-6	1. General description 2. Delivery 3. Language use 4. Topic development
IELTS	Language knowledge	interview	9	1	1. Fluency and Coherence 2. Lexical resource 3. Grammatical range and Accuracy 4. Pronunciation
ACTFL	Language use	interview	11	1	1. Global tasks and functions 2. Context/content, 3. Accuracy 4. Text types
MUET	Language knowledge	Individual presentation & group discussion	6	2	1. Task fulfilment 2. Language 3. Communicative ability

2.3 Addressing the Issue of Reliability

Analytic rating scales or analytic scoring rubric are normally used by examiners or raters to judge several components, traits or dimensions of a performance independently and each component is assigned with a matching score and descriptor. This compartmentalised oral proficiency assessment scale is generally constructed so that raters or examiners can judge each component separately rather than giving a single score for the entire performance, like in a holistic rating scale (Weigle, 2002). Even though both rating scale types share the same goal, that is to assess oral skills of learners, each rating scale differs in terms of function, application and purpose.

The practicality of analytical rating scales lies on its informative function to describe learners' strengths and weaknesses. Applying this scale in assessment requires the assessors to pay attention to each criterion listed, which may involve heavier cognitive processing and thus, requires longer processing time. In terms of purpose, analytical scales are normally used for diagnostic reason whereby the results inform subsequent teaching curricula, pedagogy or remedial tasks. One main advantage of the analytic scoring method over the holistic counterpart is that it provides a higher reliability (Goulden, 1994). Weigle (2002) also echoed the same sentiment that compared to holistic scoring, analytic scoring is more useful for second-language learners, and more reliable (Bordin Chinda, 2009). Furthermore, due to its pedagogical relevance, a criterion-referenced, analytic rating scale is more reliable than a norm-referenced, holistic rating scale (Nakatsuhara, 2007) considering that learners are not assessed based on an overall impression alone since each criterion is given due consideration. A study by Sato (2012) found that raters were able to judge both linguistic components of performance and content elaboration or development, even though specific descriptors were not provided in the study. In the study, nine raters examined 30 participants' monologues on three topics by using intuitive judgements. These raters assigned scores based on five criteria: grammatical accuracy, fluency, vocabulary range, pronunciation, and content elaboration/development. Findings revealed that content elaboration or development contributed substantively to the intuitive judgments and overall score.

3.0 THE STUDY

The study was conducted in the second semester of a one-year matriculation programme in Malaysia. A matriculation programme is a preparatory programme for students with *Sijil Pelajaran Malaysia* (SPM), equivalent to O-Levels, to qualify them for degree courses in

science, technology and the professional arts in public and private universities (www.moe.gov.my). The study involved 11 learners from an electrical engineering stream, who were randomly selected from 26 tutorial groups. The learners' average age was 18 and in terms of educational background, they were screened for the one-year matriculation programme by the Matriculation Division, Ministry of Education Malaysia based on their SPM results. The demography of the learners is as follows:

Table 2: Demography of learners

Course	Age	Male	Female	Total
Electrical engineering	18	7	4	11

3.1 Developing Learner-Driven Oral Assessment Criteria

The instrument used in this study was the group's learner-driven oral assessment criteria. These criteria were developed by learners themselves in five phases. These phases were conducted during their matriculation English subject for nearly two weeks. Learners were first encouraged to share their experience of sitting for a speaking test. They were later asked to elicit what they thought were the criteria for oral assessment. Then, they got into groups of three or four. They discussed the criteria by providing the scale and criteria. The researcher did not in any way influence their choice of criteria, levels and scores.

After producing the oral assessment criteria, each group (a group consisted of three or four members) presented their oral assessment criteria to the class. They were required to explain the choice of criteria as well as their scoring levels. Their peers were also encouraged to offer constructive feedback so that the oral assessment criteria would be comprehensible for everyone.

Next, all learners discussed the most suitable oral assessment criteria to be used by the class by integrating or combining criteria from each group. The oral assessment criteria were finalised after each group contributed their feedback. At the end of the final phase of oral assessment criteria development, a set of learner-driven oral assessment criteria was unanimously accepted and agreed to be used. The final product is shown in Table 3.

Table 3: Learner-driven oral assessment criteria

Course	Criteria Selected	Scaling structure	Labels
Electrical engineering group	1. Content:	5 rating levels	1. Very Poor
	1.1 Ideas		2. Poor
	1.2 Elaboration		3. Moderate
	1.3 Examples		4. Good
	1.4 Conclusion		5. Very Good
	2. Language:		
	2.1 Fluency		
	2.2 Grammar		
	2.3 Intonation		
	2.4 Vocabulary		
	3. Presentation:		
	3.1 Attire		
	3.2 Voice		
	3.3 Confidence		
	3.4 Facial expression		
	3.5 Body language		

3.2 Procedure

Learners in electrical engineering group had three contact hours per week with the researcher. In the first hour of the first week, learners discussed the criteria in groups of three or four. By discussing the criteria, learners may become familiar with the criteria selected. In the second hour, each group was tasked to present their oral assessment criteria to the whole class. After each presentation, their classmates would seek clarification on the assessment criteria as well as offer feedback to improve the assessment criteria. This constructive feedback was jotted down by one of the group members. In the third hour, each group's revised oral assessment criteria was displayed in the classroom and learners scrutinised each assessment criteria from each group. They also jotted down each group's strengths and weaknesses, if there is any. After everyone read all three oral assessment criteria, they returned to their group and discussed the best oral assessment criteria which they thought could be used for oral presentation assessment later on. After they had chosen the assessment criteria, the researcher elicited further

improvement from the learners. These suggestions were later incorporated into the final version of the assessment criteria for the group. The final version of the assessment criteria contained three main criteria (Content, Language and Presentation) with thirteen sub-criteria (as presented in Table 3). In peer-assessment process, learners were required to rate all thirteen sub-criteria. To test the assessment criteria, the researcher asked for three volunteers in the group in the fourth hour (first hour in the second week of instruction). These volunteers were asked to prepare a two-minute speech on the topic that they felt most comfortable speaking. To ensure similar level of difficulty for the topic, the researcher informed the volunteers that the topics must related to their daily life. While each volunteer was presenting his or her speech, their peers rated them with the finalised oral assessment criteria previously decided upon. They rated the speakers with the forms given and these forms were collected and tabulated in the Facets software. Table 4 indicates the processes in developing the learner-driven oral assessment criteria.

Table 4: Learner-driven oral assessment criteria procedure

PHASE	PROCESS	PROCEDURE
Phase 1	Discussion of Oral Assessment Criteria	• First week: first hour • Three groups • Listing possible criteria for oral assessment
Phase 2	Drafting of Oral Assessment Criteria	• First week: first hour • Three groups • Deciding the best criteria to be included in the oral assessment criteria form
Phase 3	Presentation of Oral Assessment Criteria to Peers	• First week: second hour • Three groups • Seek clarification and offering feedback
Phase 4	A Display of Finalised Assessment Criteria from Each Group to Peers	• First week: third hour • Three groups • Scrutinising assessment criteria from each group • Choosing the best set of assessment criteria • Discussing (with researchers) ways to improve ONE final set of assessment criteria
Phase 5	Testing the Oral Assessment Criteria	• Second week: first hour • Three volunteers prepare a two-minute speech on topics related to their daily life. • Peers assess speeches based on the final set of oral assessment criteria. • Researchers collect ratings and scores from every learner.
Phase 6	Analysis of Scores	• Analysis of ratings and scores using Facets software

4.0 FINDINGS

In this study, one set of learner-driven oral assessment criteria was produced from eleven learners of electrical engineering group. These criteria were tested with three speakers and peer

assessment in Phase 5. The results are presented based on the two objectives which guided the study.

4.1 The Functionality of Learner-driven Oral Assessment Criteria Scaling Structure

To understand how each category of the oral assessment criteria functioned based on the ratings given by learner-assessors, Many-Facet Rasch Measurement (MFRM) was used to analyse and demonstrate how the rating scale category functioned during peer assessment. Table 5 on rating scale statistics and Figure 2 on probability curves are presented to demonstrate oral assessment criterion functioning.

Functionality of the oral assessment criteria scaling structure was determined through the threshold measures of the Facet output. If each category is advancing by more than 1.4 logits (Linacre, 2014), it could be inferred that each category is functioning usefully in the assessment practice. Based on Table 5, it indicates that each category was functioning usefully in the group learner-driven oral assessment criteria.

Table 5: Electrical engineering group's category statistics

DATA		QUALITY CONTROL		RASCH- ANDRICH	Response Category Name
Score	Counts Used	Average Measures	Outfit MnSq	Thresholds Measure	
1	2	-1.57	0.9		Very poor
2	44	-0.45	0.9	-3.98	Poor
3	173	0.52	1.0	-1.31	Moderate
4	171	1.94	1.0	1.19	Good
5	38	3.19	1.1	4.10	Very good

To make better sense of the statistics displayed in Table 5, the Rasch model also provides graphical perspective of the rating scale functioning. Figure 2 is the probability curve obtained from Facets output file. The x-axis represents the oral presentation's criteria and it is drawn relative to the difficulty of the item. As for the y-axis, it is drawn based on the probability of observing each category of the 5-category rating scale during peer assessment practice. In short, Figure 2 illustrates that the probability of scoring a lower category and a higher category varies with ability of the learners. In Rasch, probability curves that are prominent (the peaks

are clearly visible) indicate clearly defined categories whereas probability curves that are less prominent indicate either narrowly defined categories or considerably improbable categories (Wright & Masters, 1982). Since the peaks of probability curves were clearly visible, it could be inferred that all five categories of the learner-driven oral assessment criteria were functioning usefully.

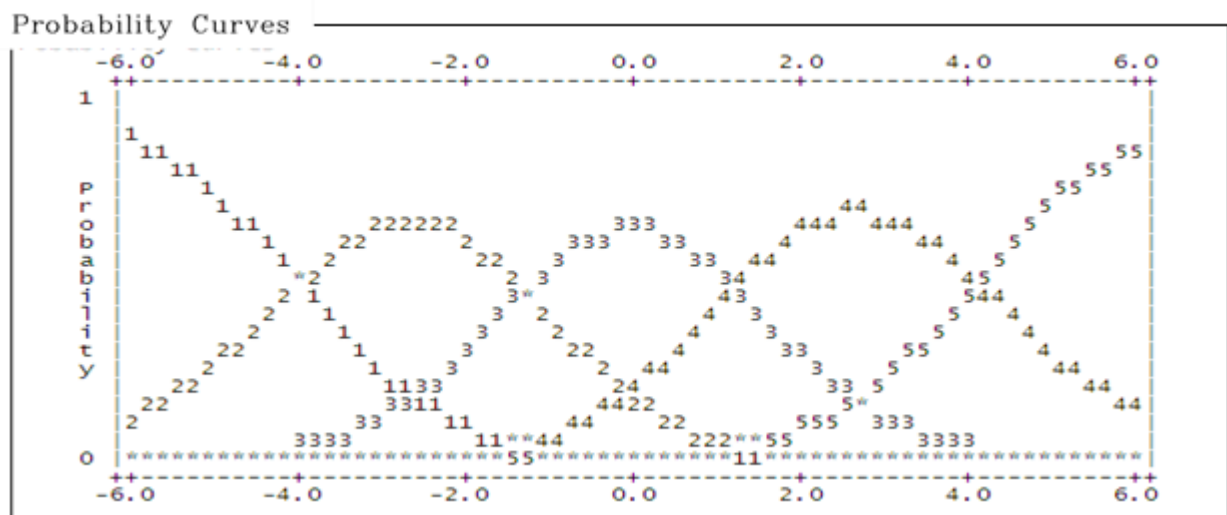


Figure 2: The electrical engineering group's rating scale structure probability curves

4.2 The Reliability of Learner-assessors in Applying Own Assessment Criteria during Oral Presentation

Since the assessment criteria were developed and produced by learners, it is pertinent that their reliability in awarding scores was gauged. The researchers would like to reiterate that this exercise was not used for summative and academic qualification purposes, it was intended as a tool for learning as prescribed in AaL. The results are displayed in Table 6.

Table 6: Learner-assessor measurement report for electrical engineering group

Participants	Obsvd Average	Fair (M) Average	Measure	Model S.E	Infit MnSq	Outfit MnSQ
E3	2.92	2.93	1.45	0.26	0.94	0.94
E6	3.05	3.06	1.11	0.26	1.46	1.46
E10	3.37	3.37	0.31	0.27	1.64	1.66
E1	3.46	3.48	0.02	0.27	0.40	0.40
E9	3.46	3.48	0.02	0.27	1.25	1.27
E4	3.49	3.51	-0.05	0.27	0.79	0.78
E2	3.56	3.59	-0.27	0.27	0.79	0.84
E7	3.59	3.61	-0.34	0.27	0.72	0.76
E5	3.72	3.74	-0.70	0.27	0.84	0.85
E11	3.72	3.74	-0.70	0.27	0.94	1.02
E8	3.77	3.80	-0.85	0.27	1.01	0.99

S.D.: 0.68; Separation 2.54; Reliability: 0.87

Table 6 shows eleven learner-assessors (E1 – E11). The three voluntary presenters who presented their speeches in Phase 5 (Testing the Oral Assessment Criteria) were E2, E9 and E11. Each presenter was rated by their ten peers during this phase. Thus, non-presenter rated all three presentations while the presenters only rated two of their peer presentations. Hence, each voluntary presenter awarded 26 ratings (13 sub-criteria x 2 presentations) while each non-presenter awarded 39 ratings (13 sub-criteria x 3 presentations). In sum, the total ratings collected from the group was 468 ratings (26 ratings x 3 voluntary presenters + 39 ratings x 10 non-presenters). From Table 6, it shows that observed average and fair average of each participant did not differ much since the data were complete for analysis. For example, E3's fair average rating was 2.93 and his observed average was 2.92. This demonstrates that even after Facet adjusted its fairness of scores for all participants, E3 was rated as the most severe learner-assessor with 1.45 logits for its severity measure.

Next to 'Fair (M) Average' column is a 'Measure' column. This column displays participants' severity in awarding ratings in Rasch logits. From the column, it shows that there were more severe learner-assessors than lenient ones. The highest measure for the most severe learner-assessors was 1.45 logits (E3) while the lowest measure for the most lenient learner-assessor was -0.85 logits (E8). The difference between the severest learner-assessor and the

most lenient learner-assessor was 2.3 logits. The difference of 2.3 logits indicates that the learner-assessor's severity in awarding rating was quite clustered.

At the bottom of the table is standard deviation (S.D.), separation and reliability. The Rasch S.D. reported 0.68 which represents the distance of each datum from the mean 0. In MFRM, it is also important to study separation. Separation for learner-assessor's severity in rating was 2.54. From the separation value, it was clear that their severity in rating had only been separated up to nearly 2 levels. In MFRM, Infit means inlier-sensitive or information-weighted fit, an indication of internal consistency of the rater or assessor. From the infit value report, E1 and E10 could be perceived as inconsistent learner-assessors. The range for productive measurement should be between 0.5 – 1.5 logits (Linacre, 2014). In MFRM, outfit means outlier-sensitive, an indication that there is a misfit of item. In this context, E1 and E10 were the misfit learner-assessors as the logits reported were below 0.5 logits and above 1.5 logits.

Reliability was reported at 0.87 and this high reliability could be attributed to the 13 sub-criteria which the learner-assessors had to rate for three presentations. Linacre (2014) stated that to obtain high reliability, one of the ways is to devise an instrument with many items. For the internal consistency of ratings, all assessors except E1 and E10 were within the range for internal consistency, meaning that they were able to maintain consistency when they were rating their peers.

5.0 DISCUSSION

In terms of rating categories, the eleven learners from electrical engineering group prescribed formal rating labels with 'very poor', 'poor', 'moderate', 'good' and 'very good'. It is also interesting to note that the electrical group did not use 'excellent' as the highest category. Perhaps they were aware of their own proficiency levels that they subconsciously did not include the label. This is consistent with some studies which claimed that learners were generally aware with their levels (Boud et al., 2013; Sato, 2012; Lim, 2007). In addition, learners seemed to understand that in developing assessment criteria, there must be a progression of ability from low to high. In addition, from the findings, it could be deduced that the rating scale was functioning usefully with five rating scales since the probability curves show that the peaks are more visible, and the categories are distinctively defined. This supports Luoma's (2004) suggestion of developing four or five categories of rating scale for raters.

Despite the threshold measure showing more than 1.4 logits between levels, the graph shows that the tendency to award mid categories (ratings 3 and 4) was prevalent amongst learners. Hence, in developing learner-driven assessment criteria, teachers should be aware of the disadvantages in having limited categories since it could affect learners' judgement to award mid-range categories. In terms of reliability, learner-assessors in the group were generally able to award ratings consistently. In fact, the existence of two misfit learner-assessors might mirror the reality as it would be quite impossible for every assessor to assess consistently in a group.

6.0 CONCLUSION

Ideally, having learners engineer their own assessment could lead to a learning culture, consistent with the establishment of learner-centred approach to learning. Yet, this has yet to reach the classroom assessment levels despite the theory of constructivism being applied to curriculum instructions. Granted, learners may not have the specific or scientific knowledge to construct and assess their own oral skills. However, the results from this small-scale study reveal that learners were able to develop a functioning rating scale to assess their oral presentation. Far more than that, most learners were able to assess reliably since they understood the assessment criteria as part of their learning process. This study may be a small step towards learner-centred approach to assessment and perhaps future research could focus on the thinking process involved in developing the criteria. In addition, a study on the effects of learner-driven assessment criteria on oral proficiency's progress could have also been conducted.

REFERENCES

- Aslan, S. & Reigeluth, C. (2015). Examining the challenges of learner-centered education. *Phi Delta Kappan*, 97(4), 63-68.
- Boud, D., Lawson, R., & Thompson, D. G. (2013). Does student engagement in self-assessment calibrate their judgement over time? *Assessment & Evaluation in Higher Education*, 28(8), 941-956.
- Bordin Chinda, M. A. (2009). *Professional development in language testing and assessment: A case study of supporting change in assessment practice in in-service EFL teachers in Thailand*. (Unpublished doctoral dissertation). University of Nottingham, United Kindom.
- Earl, L. (2013). *Assessment as learning*. Thousand Oaks, CA: Corwin Press.

- Gardner, J. (2012). *Assessment and learning* (Second Edition). London: SAGE Publications Ltd.
- Gielen, S., Dochy, F., Onghena, P., Struyven, K., & Smeets, S. (2011). Goals of peer assessment and their associated quality concepts. *Studies in Higher Education* 36(6), 719–735.
- Gikandi, J. (2011). Achieving meaningful online learning through effective formative assessment. In G. Williams, P. Statham, N. Brown, B. Cleland (Eds.), *Changing demands, changing directions* (pp.452-454). Tasmania, Australia: Proceedings Ascilite Hobart.
- Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *The Journal of Research and Development in Education*, 27(1), 73-82.
- Graue, M. E. (1993). Integrating theory and practice through instructional assessment. *Educational Assessment*, 1(4), 283-309.
- Hamidi, E. (2010). Fundamental issues in L2 classroom assessment practices. *Academic Leadership: The Online Journal*, 8(2), 1–17.
- Keppell, M. & Carless, D. (2006). Learning-oriented assessment: A technology-based case study. *Assessment in Education*, 13(2), 179–191.
- Lim, H. (2007). A study of self- and peer assessments of learners' oral proficiency. *Camling*, 1(1), 169-176.
- Linacre, J. M. (2004). Test validity and Rasch measurement: Construct, content, etc. *Rasch Measurement Transactions*, 18(1), 970-971.
- Little, D. (2005). The common European framework and the European language portfolio: Involving learners and their judgements in the assessment process. *Language Testing*, 22(3), 321–336.
- Luoma, S. (2004). *Assessing speaking*. Cambridge, U.K.: Cambridge University Press.
- Nakatsuhara, F. (2007). Developing a rating scale to assess English speaking skills of Japanese upper-secondary students. *Essex Graduate Student Papers in Language & Linguistics*, 9(1), 83–103.
- Nunan, D. (1997). Does learner strategy training make a difference? *Lenguas Modernas*, 24(1), 123-142.
- Reinders, H. (2010). Towards a classroom pedagogy for learner autonomy: A framework of independent language learning skills. *Australian Journal of Teacher Education*, 35(5), 40-55.

- Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, 29(2), 223–241.
- Shepard, A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4–14.
- Spiller, D. (2012). *Assessment matters: Self-assessment and peer assessment*. Teaching development unit. New Zealand: The University of Waikato.
- Taras, M. (2008). Issues of power and equity in two models of self-assessment. *Teaching in Higher Education*, 13(1), 81–92.
- Vickerman, P. (2009). Student perspectives on formative peer assessment: An attempt to deepen learning? *Assessment & Evaluation in Higher Education*, 34(2), 221–230.
- Weigle, S. C. (2002). *Assessing writing*. Cambridge, U.K.: Cambridge University Press.
- Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: Mesa Press.