Man In India, 96 (6) : 1801-1808

GAUGING ESL LEARNERS' CEFR RATINGS ON ORAL PROFICIENCY IN RATER TRAINING

Mardiana Idris and Mohamad Hassan Zakaria

Rater training is fundamental in reducing rater variability in self- and peer assessments practice within the paradigm of assessment *as* learning (A*a*L). Since Malaysian education system is examination-oriented in which assessment *of* learning (summative) and assessment *for* learning (formative) dominate, ESL learners are rarely asked to rate themselves or their peers as the system is still sceptical in entrusting learners with the role of assessors. Learners are normally perceived as unable to (1) assess accurately, (2) assess consistently and (3) discriminate oral proficiency components in their performance. Therefore, this study attempts to gauge learners' rating skills on these three assumptions through the use of Common European Framework of Reference (CEFR) oral assessment criteria. Quantitative analysis was conducted using the Rasch model while a short semi-structured interview was used to support the quantitative results obtained. Findings from this study suggest that ESL learners were generally able to rate accurately and consistently after rater training. The rater training had also somewhat sensitized these learners to CEFR oral proficiency components but they were still grappling in confidently rating range and accuracy. Based on these findings, it indicates that ESL learners' are ready for AaL paradigm shift and this will possibly launch a platform for self- and peer assessments practice in ESL classrooms. Such assessments will promote active learning and effective learner-centred classroom.

INTRODUCTION

In many self- and peer assessments studies on oral proficiency, rater training is fundamental in reducing rater variability as failing to conduct such training results in 'construct-irrelevant variance' (Kang, 2012). Understandably, reducing rater bias or mitigating rater effects are crucial to ensure fairness of judgments particularly when these assessments are framed within assessment *of* learning (summative) and assessment *for* learning (formative) whereby scores determine placement and certification. However, only a few rater training was reported for assessment *as* learning (A*a*L) – a recent assessment paradigm that advocates learners' involvement with assessment criteria in order to foster learners' critical thinking and independent learning. A*a*L 'focuses on the role of the student as the critical connector between assessment and their learning' (Earl, 2013: 28). In essence, learners are responsible for aligning assessment objectives with their own personal learning endavours as well as external expectations which normally come from teachers, schools, examination boards or society. Since learners are required to act as assessors of their own learning, it is central to the assessment process that learners understand

Address for communication: **Mardiana Idris**, PhD candidate, Faculty of Education, Universiti Teknologi Malaysia, Malaysia, *E-mail: anaidram7337*@hotmail.com and **Associate Professor Dr Mohamad Hassan Zakaria**, Language Academy, Universiti Teknologi Malaysia, Malaysia, *E-mail: m-zhasan@utm.my*

and know how to apply rating skills in assessing their own performances. Though this assessment type corroborates with the objectives of successful language learning, it has not been fully embraced in Malaysia, probably due to the educational system that still favours high stakes examinations (Saw, 2010; Tan, 2011) and possibly from lack of exposure. Furthermore, oral proficiency components entail complex and intertwined features such as grammar, vocabulary, pronunciation and fluency, to name a few, which demand more instructional hours than most schools can offer. Beyond that, assessing oral proficiency accurately requires heavier cognitive load for assessors as they first have to listen, comprehend, gauge the performance by matching the speakers' level to the assessment criteria and finally, award appropriate and fair scores or bands. These simultaneous processes which happen in real time while the speaker is speaking require undivided attention and more importantly, rating skills.

Since Malaysian education system is examination-oriented, learners are rarely asked to rate themselves or their peers as the system is somewhat skeptical in entrusting learners with the role of assessors. Learners are normally perceived as unable to (1) assess accurately, (2) assess consistently and (3) discriminate oral proficiency components in their performance. Therefore, this study attempted to gauge learners' rating skills based on Common European Framework of Reference (CEFR) oral assessment scale. This scale will be elaborated further in the next section. Three research questions which framed the study were:

1. Did ESL learners apply the CEFR scale accurately in rater training?
2. Did ESL learners rate consistently in rater training?
3. To what extent did learner rater training sensitize ESL learners' to CEFR oral proficiency components, namely overall impression, range, accuracy, fluency and coherence?

DESCRIPTIONS OF THE STUDY

Common European Framework of Reference (CEFR)

Common European Framework of Reference (CEFR) is an empirically developed measurement for listening, speaking, reading and writing. However, in this study, only speaking scales were used as the instrument to elicit learners' ratings in assessing three speakers, featured in Council of Europe (CoE) videos. The ratings used in this study were based on the reconstructed CEFR oral assessment criteria (CoE, 2001), consisting of statements that described the language learners' performance at six levels: A1, A2, B1, B2, C1, and C2 as the highest level. CEFR was used in this study as it was developed empirically, using the views of practicing teachers. Though it may seem that it suits only European learners, a few studies showed that it operated well in different regions and different learners of the world (Glover, 2011). Since the task excluded pair format, the components adapted were

only five criteria: (1) overall impression, (2) range, (3) accuracy, (4) fluency, and (5) coherence. Interaction was not used in order to eliminate interlocutor effect (Fulcher, 2010; Davis, 2009). In this study, overall impression refers to holistic ratings of the speakers based on the global scale. Range reflects learners' ability to formulate ideas by using differing linguistic forms. Accuracy reflects on learners' ability to maintain consistent use of complex grammar while fluency deals with their ability to express ideas spontaneously and speaks in natural flow. As for coherence, it shows how speakers are able to structure their sentences with appropriate sequence connectors.

Reconstruction activity of CEFR oral assessment criteria grid

Participants were grouped into three or four and each group received global and analytic assessment scales of CEFR with missing descriptors. Then, jigsaw activity began in which each group discussed and filled in the missing descriptors (these missing descriptors were only on overall impression, range, accuracy, fluency and coherence). Then, the researchers discussed the descriptors with participants, steering them towards placing the descriptors correctly into the grid. After these descriptors were correctly placed, the complete scale was given to each participant. The rational for jigsaw activity was to sensitize them to the descriptors before they viewed the videos (Ibberson, 2012).

METHODOLOGY

Participants

In this study, 10 Band 3 MUET (Malaysian University English Test) learners were purposely selected. Logistically, it was impossible to randomly select participants due to learner, teacher and school constraints. Though the number was small, they represented the majority of ESL learners who normally scored Band 3 in MUET examinations (Malaysians Examinations Council, 2010). MUET is an English proficiency test of the four language skills, administered to tertiary level learners by Malaysian Examination Council. The results are depicted in aggregated score that categorizes the learners into six bands, ranging from Band 1 (extremely limited user) to Band 6 (very good user). Only participants who scored Band 3 (modest user of English) were selected for this study as they could understand the CEFR descriptors as well as provide more responses for the interview. 6 females and 4 males volunteered with 5 of them Malay, 4 Chinese and 1 Indian, mirroring the three main races in Malaysia. All of them were 18 years old.

DVD of spoken performances illustrating CEFR levels

After reconstruction activity of the CEFR oral assessment criteria, the participants viewed three DVDs of levels A1 (Tifaine: T), B2 (Paul: P) and C2 (Xavier: X) of

CEFR spoken performance. They were required to rate the speakers by filling in the CEFR rating forms (Appendix 1) based on the CEFR criteria previously reconstructed. These DVDs were viewed on CoE website (http://www.coe.int/) whereby the calibrated examples were available for training purposes from http://www.ciep.fr/. Ratings, comments and transcriptions of the calibrated examples were also downloaded to guide and assist rater training.

Quantitative analyses of ratings

Participants' accuracy and consistency in rating the speakers were computed using Rasch measurement model (Bond and Fox, 2007) by means of Winstep computer programme (Version 3.72.3). Although the sample size was small, it was adequate for analysis in Winstep as it fulfilled the minimum 10 observations per category (Linacre, 2014).

Qualitative analysis of interviews

A brief semi-structured interview on rater training were conducted to explore participants' (a) involvement with oral assessment criteria, (b) understanding of CEFR oral assessment descriptors, and (c) general view on rater training practice.

RESULTS

In Rasch analysis, testing the fit between data and the model was conducted through quality control of fit statistics with z-scores (or Zstd), mean-square (MnSq), infit and outfit (Linacre, 2014). The value of Zstd was 0.0, indicating that the data fit the model. The MnSq value for infit and outfit were 1.0 and 0.98 respectively which indicated that the measurement was accurate. The subsequent results were reported according to the research questions.

RQ1: 1. Did ESL learners apply the CEFR scale accurately in rater training?

Accuracy in this study was viewed within the context of how learners' ratings matched the experts. The third column (Exact Obs%) of Table 1 shows percentage of ratings that matched the experts. Only five learners (L03, L10, L01, L02 and L05) achieved agreement more than the acceptable percentage (70%) while L04 scored less than 50% agreement with the CEFR experts. This could be due to L04 ratings of Xavier (C2 speaker), who rated him as C1 during rater training. However, L04 ratings were only one point lower than expected but were still within proficient speakers' range (C1 and C2) according to CEFR oral assessment scale. The last column (Match Exp%) shows the agreement percentage that was expected if the data fit the model perfectly. L03 met the expectation of the model with 100% while L06, L07, L08 and L09 did not meet the expectation by only 2.6%. This indicates that the four learners' ratings were more random than the model predicts.

However, L01, L02 and L05 observed percentages show 10.7% higher than the expected model, an indication that the ratings were predictable.

TABLE 1: LEARNERS' RATING AGREEMENT WITH CEFR EXPERTS

Learner	*Total Score*	*Exact Obs%*	*Match Exp %*
L03	15	100.0	100
L10	14	93.3	93.4
L01	10	80.0	69.3
L02	10	80.0	69.3
L05	10	80.0	69.3
L06	10	66.7	69.3
L07	10	66.7	69.3
L08	10	66.7	69.3
L09	10	66.7	69.3
L04	9	46.7	66.1

RQ2: Did ESL learners rate consistently in rater training?

Learners' internal consistency of ratings was gauged through infit and outfit mean square residuals (MSq). Ideally, both should be close to one within the range of 0.5 to 1.5 (Linacre, 2014). Outfit MSq is sensitive to extremely unexpected individual ratings (outliers) while infit MsQ is less sensitive to outliers but more sensitive to unexpected rating patterns – an indication of internal consistency (Yan, 2014). In Table 2, only L04 appeared to have infit MsQ higher than 1.5, suggesting a tendency to rate inconsistently and unpredictably. However, infit and outfit statistics of nine other learners were within the range, indicating that almost all learners were largely internally consistent. Since L03 scored 100% agreement with the experts, the infit and outfit statistics reported 'maximum measure' as the learner was behaving like a 'rating machine', thus no longer included in the measurement situation. These results suggest that most learners were able to rate consistently and rater training could have contributed to the consistency. This is consistent with many studies which found that rater training may become a mechanism to

TABLE 2: INTERNAL CONSISTENCY OF LEARNERS' RATINGS

Learner	*Model SE*	*Infit MSq*	*Outfit MSq*	*PtBis Corr*
L03	1.83	Max. measure	Max. measure	0
L10	1.05	1.07	1.03	0.07
L01	0.58	0.82	0.77	0.56
L02	0.58	0.90	0.84	0.46
L05	0.58	0.66	0.59	0.77
L06	0.58	0.82	0.72	0.58
L07	0.58	0.82	0.72	0.58
L08	0.58	1.22	1.20	0.03
L09	0.58	1.22	1.20	0.03
L04	0.56	1.54	1.77	-0.46

eliminate rater error and improve consistency (Farrokhi, Esfandiari and Schaefer, 2012; Carey, Mannell and Dunn, 2010).

RQ3: To what extend did learner rater training sensitize ESL learners' to oral proficiency components, namely overall impression, range, accuracy, fluency and coherence?

In relation to their involvement with rater training, many learners repeatedly used the word 'helped' and 'improve' during their interview.

L03: I believe my oral skills will improve if I get involved with rater training.

L01: With rater training, it helped me to understand oral assessment criteria.

Although they acknowledged that rater training was helpful in improving their oral skills, more than half requested for longer training session.

In terms of understanding CEFR oral assessment descriptors, majority felt that the descriptors were easily understood and all of them concurred that they were able to categorize the speakers and themselves according to the CEFR levels on overall impression, fluency and coherence. However, only some learners were confident in rating range and accuracy. All learners agreed that the descriptors helped them in identifying their strengths and weaknesses.

Generally, learners displayed positive disposition towards rater training practice as evident from the responses given by L10, L04 and L05(verbatim):

L10: This training helps me to understand my oral abilities; I know what I need to improve about my speaking.

L04: I think this oral assessment criteria will bring a lot of advantages if me involve in this assessment criteria although it may made me feel difficult because my English level is really low and poor.

L05: I think that the oral assessment criteria helped me to know how to differentiate each level according to accuracy, fluency, range and coherence.

DISCUSSIONS AND CONCLUSIONS

Generally, findings from this study were consistent with most literature on rater training, rater behaviour and oral proficiency but the difference was situated in the assessment paradigm in which this study was framed within A*a*L context while most studies were centered on A*o*L and A*f*L.

Findings on learners' accuracy in rating the CEFR speakers suggest that most of them, though they were not as experienced as teachers or trained raters, were able to match the ratings awarded by CEFR experts. This may indicate that Malaysian ESL learners are somewhat ready for A*a*L practice in the classroom. We were cautious in our claim as studies have also shown that raters' ratings are 'somewhat idiosyncratic' (Yan, 2014) even after repeated training while learners in this study were trained only once. Hence, learners need to be prepped sufficiently so that they become confident in awarding accurate ratings as learners may have

been conditioned to trust others' judgements than their own in their current assessment context. In training the learners, it is important to develop learners' rating skills in A*a*L context rather than gearing the learners to operate as rating machines as envisioned in AoL or A*f*L systems.

Results on consistency of ratings in this study were similar to findings reported in studies on rating practice (Saito, 2008; Lim, 2011; Elder *et.al*., 2005) Though the practice was short (1 hour and 30 minutes), most learners were able to rate consistently based on the infit mean-square measures reported. Though this may sound promising for A*a*L to be established as classroom practice, we would also like to caution that participants in the study were modest users of English (Band 3 MUET) who may understood majority of the words used in the criteria which subsequently translate to their ability to rate the speakers based on the descriptors listed. It may produce different results should different set of participants is used such as limited users of English or native speakers.

Qualitative findings of this study complement the analysis of ratings awarded by participants. In order for them to apply the criteria, they must first understand its descriptors. It was anticipated that learners will find the criteria difficult and confusing as it was their first time involvement with the activity. Though some learners displayed positive outcome from the activity when they learnt how to rate, it was also interesting to notice that some learners rated the speakers with criteria that were not included in the assessment criteria. For example, a few participants mentioned that CEFR speakers' pronunciation in the videos affected their comprehension when in the CEFR oral assessment criteria, pronunciation is not listed. A few studies mentioned this tendency of raters who had different set of criteria when rating even when rating practice was provided (Eckes, 2005).

In conclusion, findings from this study suggest that ESL learners were generally able to rate accurately and consistently after rater training. The rater training had also somewhat sensitized these learners to CEFR oral proficiency components but they were still grappling in confidently rating range and accuracy. Based on these findings, it indicates that ESL learners' are somewhat ready for A*a*L paradigm shift and this will possibly launch a platform for self- and peer assessments practice in ESL classrooms. Such assessments will promote active learning and effective learner-centred classroom. In order to improve accuracy of learners' ratings, better results would have been achieved if videos of A2, B2 and B1 were shown. In fact, future studies might want to use larger sample of ESL learners, longer period of rater training, different oral assessment scales or different oral proficiency components.

References

Bond, T.G & Fox, C.M (2007). Applying the Rasch Model:Fundamental measurement in the human sciences (Second Edition). London: Lawrence Erlbaum Associates.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2010). 'Does a Rater's Familiarity with a Candidate's Pronunciation Affect the Rating in Oral Proficiency Interviews?', *Language Testing,* 28(2): 201–219.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge, UK: Cambridge University Press/Council of Europe.

Davis, L. (2009). 'The Influence of Interlocutor Proficiency in a Paired Oral Assessment', *Language Testing,* 26(3): 367–396.

Earl, L.M. (2013). Assessment as Learning: Using Classroom Assessment to Maximize Student Learning (Second Edition). California: CORWIN

Eckes, T. (2005). 'Examining Rater Effects in Testdaf Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis', *Language Assessment Quarterly,* 2(3): 197-221

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). 'Individual Feedback to Enhance Rater Training: Does It Work?', *Language Assessment Quarterly,* 2(3): 175–196.

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). 'A Many-Facet Rasch Measurement of Differential Rater Severity / Leniency and Teacher Assessment', *Journal of Basic and Applied Scientific Research,* 2(9): 8786–8798.

Fulcher, G. (2010). 'The Reification of the Common European Framework of Reference (CEFR) and Effect-Driven Testing', Advances in Research on Language Acquisition and Teaching, 15–26.

Glover, P. (2011). 'Using CEFR Level Descriptors to Raise University Students' Awareness of Their Speaking Skills', *Language Awareness,* 20(2): 121–133.

Ibberson, H. (2012). An investigation of non-native learners ' self-assessment of the speaking skill and their attitude towards self-assessment. Doctor of Philosophy, University of Essex.

Kang, O. (2012). 'Impact of Rater Characteristics and Prosodic Features of Speaker Accentedness on Ratings of International Teaching Assistants' Oral Performance', Language Assessment Quarterly, 9(3): 249–269.

Lim, G. S. (2011). 'The Development and Maintenance of Rating Quality in Performance Writing Assessment: A Longitudinal Study of New and Experienced Raters', Language Testing, 28(4): 543–560

Linacre, J. M. (2012). Facets (Version 3.70). Beaverton, OR: Winsteps.com. Retrieved from *http:// www.winsteps.com/facets.htm*

Malaysian Examinations Council (2010). Laporan Kajian Pencapaian Malaysian University English Test (MUET) 2002 - 2006. Selangor: Majlis Peperiksaan Malaysia.

Saito, H. (2008). 'EFL Classroom Peer Assessment: Training Effects on Rating and Commenting', *Language Testing,* 25(4): 553–581.

Saw, L. O. (2010). 'Assessment Profile of Malaysia: High-Stakes External Examinations Dominate', Assessment in Education: Principles, Policy & Practice, 17(1): 91–103.

Tan, Y. S. (2011). 'Democratization of Secondary Education in Malaysia: Attitudes Towards Schooling and Educational Aspirations', *Asia Pacific Journal of Education,* 31(1): 1–18.

Yan, X. (2014). 'An Examination of Rater Performance on a Local Oral English Proficiency Test: A Mixed-Methods Approach', Language Testing, 1–27.