Full paper

Sains Humanika

# The CEFR Rating Scale Functioning: An Empirical Study on Self- and Peer Assessments

Mardiana Idris[a], Abdul Halim Abdul Raof[b*]

[a]*Faculty of Education, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia*
[b]*Language Academy, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru Johor, Malaysia*

*Corresponding author: m-halim@utm.my

**Abstract**

One of the criticisms on the Common European Framework of Reference (CEFR) rating scales pertains to the lack of reference to the performance of learners in the construction process of the scales. Therefore, this study attempted to delve into rating scale functioning by English as a Second Language (ESL) learners during self-assessment and peer assessment of their oral proficiency practice. Two objectives guided the study: 1) to gauge the overall rating scale functioning and 2) to measure each criterion scaling structure. Three self- and peer assessments' cycles were conducted in three months. In each cycle, eleven learners recorded their own speech, uploaded their video clips to a private YouTube channel and assessed their own videos as well as selected peers based on five CEFR oral assessment criteria with six levels of ratings (A1-C2). Findings revealed that four of the CEFR levels were utilised (B1-C2). Categories A1 and A2 (basic user level) however, were not observed during the practice. Analysis from the Many-Facet Rasch Measurement (MFRM) indicated that utilised categories seemed to function usefully since each category observed was advancing by more than 1.4 logits. Category B2 dominated four criteria of ratings awarded while B1 dominated the rating distribution for fluency. The implications of this study will be discussed in relation to rating scale development, specifically on matching learners' proficiency to the psychometrically developed rating criteria as well as illustrating assessment as learning approach in the ESL classroom where learners become the key assessors for their own performance.

*Keywords*: Common European Framework of Reference (CEFR), self-assessment, peer assessment, rating scale functioning



## 1.0 INTRODUCTION

Since its inception in 2001, the Common European Framework of Reference (CEFR) has been robustly applied to the field of language assessment and testing (Little, 2007), which somewhat clouds the initial purpose of the CEFR, that is to harmonise language, teaching and assessment within the use of its framework (Council of Europe, 2001). With its six-level empirically developed rating scales (Jones & Saville, 2009), researchers were drawn to use it as instruments for self- and peer assessments (SAPA) in measuring oral proficiency (Glover, 2011;Ibberson, 2012; Hulstijn *et. al.*, 2011). Although the scale was empirically derived, assessment theorists argued that it was not based on performance data since there was no reference to the performance of learners or test takers on specific tasks, or even perceptions of the value of performances (Fulcher, Davidson, & Kemp, 2011). With this argument in mind, the main purpose of this study was to gauge the extent of the CEFR rating scale functioning during the SAPA practice on the ESL learners' oral proficiency. Two objectives guided this study which were 1) to gauge the CEFR rating scale functioning during SAPA practice, and 2) to measure each criterion scaling structure used in this study which were overall impression, range, accuracy, fluency and coherence.

In reporting the study, this paper is structured into five sections. The first section reviews relevant literature pertaining to the CEFR rating scale development and the SAPA practice. These reviews are the basis for the proposed two research questions. The second section describes and justifies the method and procedure used. The third section reports the results of the study. Discussion of findings is in the fourth section and the final section concludes as well as presents the implications of this study in relation to rating scale development and assessment *a*s learning (A*a*L) in the ESL classroom.

## 2.0 THE CEFR RATING SCALE AND SAPA PRACTICE

The CEFR has been widely successful beyond Europe due to two factors: the first is related to its perceived useful six-level labels (A1, A2, B1, B2, C1 and C2) and secondly, these levels are empirically developed and validated, which subsequently affect 'the drafting of objectives, targets, and outcomes in language learning programs in different contexts for different uses and purposes' (Figueras, 2012: p. 479). Consequently, the CEFR has also been applied for SAPA practice on listening and reading (Alderson, Figueras, & Kuijper, 2006) as well as writing (Huhta *et. al.*, 2014). North (2014) remarked that the CEFR should be perceived as a starting point for further development of a rating scale as well as for the purpose of accommodating a test's context and requirements. Unfortunately, the researchers found only limited studies on how the CEFR rating scale functioned in each test or assessment. Thus, the next section will review relevant literature on the CEFR rating scale development and SAPA practice.

## The CEFR Rating Scale Development: The History

The CEFR was first published in two draft versions in 1996 and the document was later revised and published in French and English in 2001(Little, 2006). The impact of the CEFR was immediately apparent for after the German translation of the document, more than 21 countries translated it into their own languages. In 2005, a survey by the Language Policy Division on 111 respondents of 39 countries found that the most frequently used component of the CEFR was the common reference levels of the rating scales. As a result, these rating scales have become the 'common currency' in many countries in Europe and beyond its border.

However, criticisms are mounting towards the CEFR rating scales as it was not based on theories of second language acquisition (Hulstijn, 2007) or 'a principled analysis of language use within a range of domains' (Fulcher, 2010).Thus, it raises the question, 'Where does a curriculum theory stand in the CEFR rating scale development?', particularly when the scale has been referred to in a variety of state and regional official documents. The traditional curriculum theory (Tyler, 1949) generally focuses on three shared features. First, curriculum theory is institutional in orientation whereby the curricular for the school system is developed to cater to the needs of society. Second, curriculum development is largely technological and rationalistic in which it is characterised by a series of techniques, as well as employment of models and frameworks. Third, curriculum theory is orientated towards reform so as to respond to the needs of students, society and culture. However, these essential features are somewhat vague in the developmental stages of the CEFR. Valax (2011)echoed the same view and discussed in length some of the basic principles of the CEFR which were not anchored to any particular curriculum the oryand its implications to curriculum designers. Despite the needs for a curriculum theory, Schoenfeld (2016)in his hermeneutic exercise of analysing 50 nominated American Educational Research Association (AERA) presidential papers on significant changes in curricular development from 1916 to 2016,observed that there was no mention of 'curricular and philosophical giants, such as Dewey, Brownell, and Judd' (page 110) in any of the papers he reviewed. The criticisms on the absence of a curriculum theory were acknowledged by North (2014). However, he argued that the CEFR levels were scaled using the Many-Facet Rasch Measurement (MFRM) which calibrated the teachers' judgement with the scales' difficulty level. Therefore, since the formulation of the CEFR scales was based on psychometric properties, North claimed that it was entirely based on the theory of measurement(Hand, 1996). The CEFR scales were developed in repeated cycles of 4 phases and 12 steps (Fulcher, 2003) and these are illustrated in Figure 1 below.

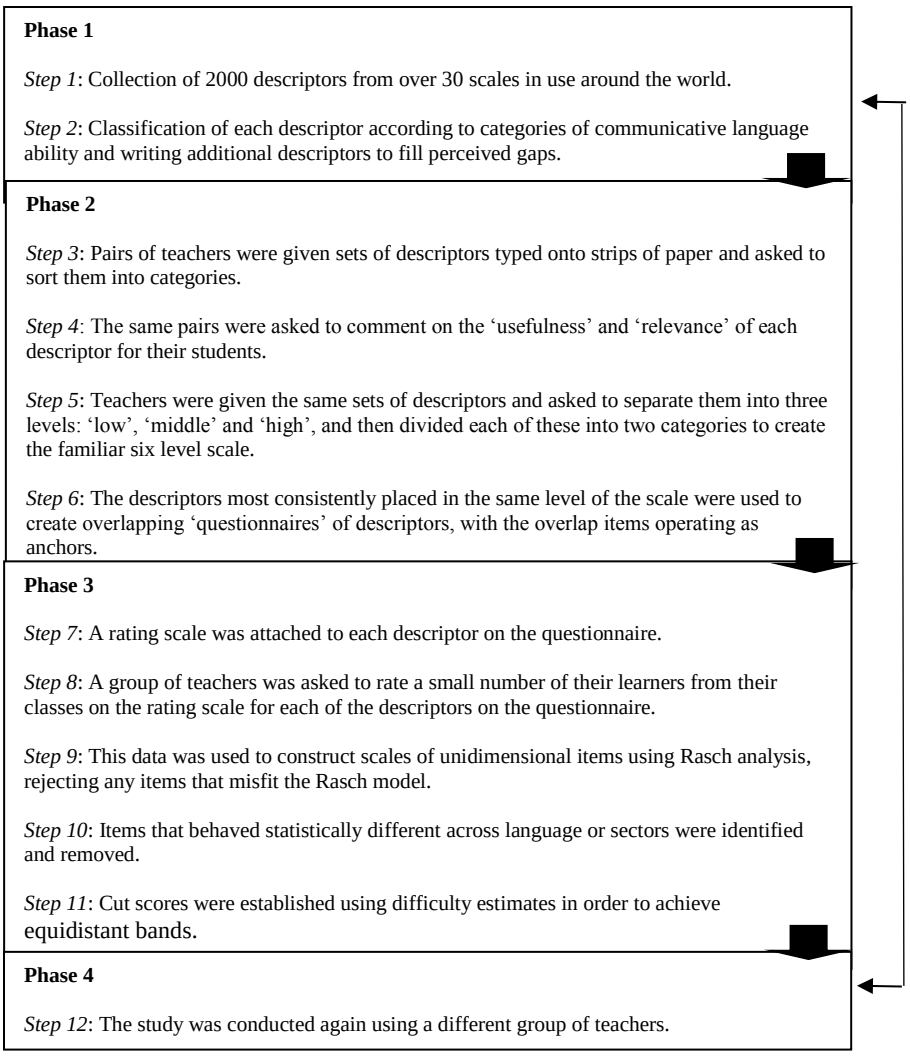


**Figure 1** The development process of the CEFR scales (Fulcher, 2003:107-113)

As illustrated in Figure 1, to calibrate the scale whereby the difficulty level of items was relatively corresponding with the teachers' judgement, the Rasch model was employed to analyse the psychometric properties of the scaling structure. Rasch analysis is advantageous as it provides sample-free and scale-free measurement. This means that the scaling does not depend on samples or tests used in the analysis. Due to these principles, the Rasch model has been used as means of analysis in many SAPA practice studies. The following section will review SAPA practicethat utilises the CEFR rating scales.

### SAPA Practice and the CEFR

Although there are studies on the effect of self-assesment on learning since 1978 (Brantmeier, Vanderplank, & Strube, 2012) and a landmark-seminal by Falchikov (1986) on influence of peer assesment on learning and organization (Topping, 1998), studies which utilised the CEFR as its instrument for assessment purpose are still developing and limited since the CEFR was introduced only fifteen years ago. In addition, studies on SAPA practice were generally conducted in the context of assessment of and for learning (summative and formative assessments) and only a few studies reported SAPA practice within assessment as learning (Nulty, 2011). One significant factor which distinguishes assessment as learning (A*a*L) from other types of assessment is the role of assessor whereby the key assessor in A*a*L is the learners themselves (Earl, 2003) which causes two concerns, i.e. are learners able to assess their own performance and are they able to use the rating scale usefully? Most studies reported favourable and advantageous effects of the CEFR in classroom teaching and language learning (Deygers & Van Gorp, 2015). However, these studies employed experienced raters and only limited studies used novice or learner-raters. This sparked the researchers' interest into gauging how the rating scale functioned with modest ESL learners who had zero experience with SAPA practice. Therefore, to guide this study, the researchers posed the following research questions:

1. To what extent does the CEFR oral assessment criteria's rating scale function during SAPA practice?
2. To what extent does each category of the CEFR oral assessment criteria function during SAPA practice?

## ■3.0 METHOD AND PROCEDURE

### Participants

Non-probability sampling, specifically purposive sampling strategy that utilised homogenous sampling, was used to select samples (Lodico, Spaulding &Voegtle, 2006) for this study. Ideally, random (or probability sampling) would have provided generalisability of the findings. However, due to practicality of study and cost consideration, there were evidence that 'much research gets done with various forms of probabilistic sampling'(Remler& Van Ryzin, 2015). Thus, the population parameter was confined to learners who studied in public schools in an urban setting. The reason for this parameter setting was to ensure that the researchers could obtain sufficient sample for this study. Table 1 shows the background of the participants.

**Table 1** Background of participants

| | |
|---|---|
| Number of participants | *Eleven* |
| Average age | *Nineteen* |
| Formal exposure to English in school | *Eleven years* |
| Academic streaming | *Science stream* |
| Level of English proficiency | *Modest user of English* |
| Exposure to SAPA practice | *Zero exposure and experience* |

### Instrument

The CEFR oral assessment criteria was selected for SAPA practice since its use for classroom assessment is 'flexible and context-amenable'. In short, its flexibility and scaling structure allow the researchers to adapt the assessment criteria to the pedagogic culture and context of use. In addition, the assessment criteria are accessible and free as the handbook and samples are easily downloaded from the council's official website.

In adapting the oral assessment criteria to the purpose of this study, only the overall descriptor of oral production (holistic rating) and the four oral assessment criteria (analytic rating) were utilised. Although interaction and its corresponding descriptors were provided in the framework, this was excluded from the oral assessment scale used in this study since the researchers wanted to control for interlocutor effects. Therefore, in this study, participants used five criteria during their SAPA practice: overall impression, range, accuracy, fluency and coherence. During SAPA practice, participants were provided with laminated oral assessment criteria which was specifically designated to the participant as the assessor for the practice. As for the assessee, it was predetermined before SAPA practice began based on the judging plan required for Many-Facet Rasch Measurement (MFRM) analysis.

### Procedure

This study was conducted for thirteen weeks. After identifying the participants who agreed to volunteer for this study, the researchers conducted a rater training session. During rater training, the researchers discussed the criteria with the participants and viewed the sample DVDs (Digital Versatile Discs) on the CEFR website (http://www.ciep.fr/en). Any discrepancies in awarding the ratings to the DVD speakers were discussed thoroughly in order to reduce rater variability during SAPA practice (Kang, 2012;Idris & Zakaria, 2016). Then, participants

rated six DVDs individually and their 330 ratings (11 participants x 6 speakers x 5 criteria) were collected and analysed using Winstep programme.

During SAPA practice, each participant recorded their own two-minute speech and uploaded their videos to a private YouTube channel. After each upload, participants rated their own videos as well as their peers based on the five CEFR oral assessment criteria. While assessing the videos, participants utilised the six-level rating scales (A1-C2). In this study, participants uploaded their videos once a month for three months. The researchers collected 990 ratings at the end of the third month of SAPA practice and analysed it using Facets programme.

## Data Analysis

In analysing the CEFR ratings awarded by participants in the private YouTube channel, the researchers decided to use the MFRM as it has been used in many performance assessment and paired comparison studies (Linacre, 2014) as well as SAPA practice. In view of this advantage, this software was deemed suitable in gauging rating scale functioning which corresponded with research questions 1 and 2 for this study.

Prior to SAPA practice and analysis of the CEFR ratings in Facets programme, several requirements had to be fulfilled first in order for the analysis to run smoothly. The first requirement was to prepare a judging plan (a rating design). Since in this study each participant was involved in self-assessment (SA)and each participant also had to conduct peer-assessment (PA)for predetermined peers, the judging plan was crucial so that there would be enough linkage between all facets and 'all parameters could be estimated without indeterminacy within one frame of reference' (Linacre, 2014). Estimation could be obtained as long as connection could be established between assessee, rater and item. Linacre (2014) proposed three judging plans while Eckes (2011) presented five rating designs. Although complete or fully cross designs lead to 'highest precision of model parameter', it is rarely practical in real assessment situations due to financial and time factors. Therefore, the researchers designed a judging plan which was similar to a rotating test book judging plan (Linacre, 2014)and Rating design B (a connected /linked design) (Eckes, 2009). The judging plan is presented in Table 2 below.

**Table 2** Judging plan for participants during SAPA practice

| | ***S1*** | ***S2*** | ***S3*** | ***S4*** | ***S5*** | ***S6*** | ***S7*** | ***S8*** | ***S9*** | ***S10*** | ***S11*** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***S1*** | S | | | | | | | | X | X | X |
| ***S2*** | X | S | | | | | | | | X | X |
| ***S3*** | X | X | S | | | | | | | | X |
| ***S4*** | X | X | X | S | | | | | | | |
| ***S5*** | X | X | X | X | S | | | | | | |
| ***S6*** | X | X | X | X | X | S | | | | | |
| ***S7*** | | X | X | X | X | X | S | | | | |
| ***S8*** | | | X | X | X | X | X | S | | | |
| ***S9*** | | | | X | X | X | X | X | S | | |
| ***S10*** | | | | | X | X | X | X | X | S | |
| ***S11*** | | | | | | X | X | X | X | X | S |

Notes:

S = Self-assessment

X = Peer assessment

## ■4.0 RESULTS

The results are reported based on two research questions previously posed using statistical reports obtained from Facets programme. Prior to reporting these primary results, output from Winstep programme in analysing participants' rating behaviour during rater training showed that their infit and outfit mean squares were within acceptable range (0.5 – 1.5 logits). This indicates that participants in this study were generally able to assess reliably and accurately during rater training and therefore, the results for the CEFR rating scale functioning and the CEFR oral assessment criteria functioning in the subsequent sections somewhat reflect the participants' application of the scale during SAPA practice.

## The CEFR Rating Scale Functioning

Table 3 shows the six-category rating scale statistics of the CEFR oral assessment criteria which were from A1 to C2 as indicated in the column 'Response Category Name'. Based on the table, it shows that A1 and A2 were not used at all during SAPA practice. As for category frequency in the 'Used' column, it shows that categories 3 (B1), 4 (B2) and 5 (C1) had large frequency counts compared to category 6 (C2). This implies that participants were able to discriminate 4 levels of oral proficiency based on the CEFR oral assessment criteria. The absence of categories 1 (A1) and 2 (A2) in the rating scale statistics was anticipated by the researchers since all the participants were categorised as modest users of English. Moreover, participants' results of rater training had shown acceptable level of reliability and accuracy which may have suggested that these participants were exercising their ratings skills as intended.

**Table 3** The CEFR six-category rating scale statistics

| DATA | | QUALITY CONTROL | | RASCH-ANDRICH | Response Category |
|---|---|---|---|---|---|
| Score | Counts Used | Average Measures | Outfit MnSq | Thresholds Measure | Name |
| 1 | 0 | | | | A1 |
| 2 | 0 | | | | A2 |
| 3 | 282 | -2.73 | 1.1 | | B1 |
| 4 | 417 | -1.71 | 0.9 | -2.66 | B2 |
| 5 | 216 | 0.88 | 0.9 | 0.19 | C1 |
| 6 | 75 | 1.87 | 1.1 | 2.47 | C2 |

Although the frequency counts for category 6 (C2) was only 75 and the participants did not utilise categories 1 (A1) and 2 (A2), evidence of good fit is shown in the Average Measure column. The average measure for category 3 (B1) was -2.73 logits, followed by -1.71 logits for category 4 (B2). Subsequently, categories 5 (C1) and 6 (C2) were also advancing together with 0.88 logits and 1.87 logits. Therefore, despite absence of category 1 (A1) and 2 (A2) from the rating scale frequency counts, there was evidence of good fit in illustrating how participants used the rating scale usefully for the four categories.

Apart from the frequency counts and average measures, validity of categorisation was gauged from the outfit mean-squares values obtained from the table. From Table 3, it shows that all outfit mean-squares were closer to 1.0, which was the expectation in the MFRM analysis. Should the mean-squares exceed considerably, it symptomises that the functioning of the rating has gone seriously wrong. With outfit mean-squares between 0.9 and 1.1, it indicates that the rating scale for these four categories was functioning usefully in SAPA practice. In order to indicate well-defined categories, Rasch-Andrich Thresholds measure was investigated. All four utilised categories were advancing by more than 1.4 logits which in turn indicates that the categories were generally well-defined.

However, Table 3 only displays how the CEFR oral assessment criteria rating scales were functioning in the assessment of SAPA participants' oral performance. It does not describe how each category in the CEFR oral assessment criteria, namely overall performance, range, accuracy, fluency and coherence functioned during SAPA practice. Hence, the following section reports the results of the second research question.

## The CEFR Oral Assessment Criteria Functioning

In order to understand how each criterion functioned based on the ratings given by participants, the Rasch partial credit model was used to analyse and demonstrate how each of the CEFR rating scale category functioned during SAPA practice. Table 4 displays the rating scale statistics of the CEFR oral assessment criteria functioning on overall impression, range, accuracy, fluency and coherence.

In order to control for the quality of measurement, average measure and outfit mean-squares have to be scrutinised first. Outfit mean-squares for all the criteria were within productive range of measurement as the range was between 0.5 and 1.5 logits. Therefore, it suggests that outliers were not detected.

From Table 4, it is evident that A1 and A2 were not used by the participants for all the CEFR criteria used in the study. Participants only utilised four categories of intermediate (B1 and B2) and advanced (C1 and C2) levels. Despite absence of beginner level categories (A1 and A2) observed, there seems to be evidence of good fit of the rating scale through the average measures reported. From the average measure of each criterion, it is evident that each measure was advancing monotonically from category 3 (B1) to category 6 (C2). For example, the average measure for overall impression was -2.69 logits for category 3 and it advanced to -1.65 logits in category 4, 1.03 logits in category 5 and finally 2.34 logits in category 6. This indicates that only these four categories were functioning properly as observations in higher rating categories produced higher logit measures. In addition, the frequency counts were distributed only across four categories, albeit unevenly. Generally, more than 40% of the rating distribution centred on category 4 (B2) for overall impression, range, accuracy and coherence. However, category 3 (B1) dominated the observation for fluency. This implies that participants were inclined to award B2 for most of the criteria observed. Although participants were categorised as modest ESL learners, between 5% and 10% of category 6 (C2) were awarded during SAPA practice for all criteria.

In order to indicate well-defined categories, the Rasch-Andrich Threshold measures were expected to advance by at least 1.4 logits. From the table, it shows that all four categories utilised were advancing by more than 1.4 logits which indicates that the categories were generally well-defined. Since these categories did not advance by more than 5 logits, it suggests that the categories were not too wide or less informative.

**Table 4** Rating scale statistics for the CEFR oral assessment criteria

| Criteria | Score | Data Counts | % | Quality Control | | Rasch Andrich Threshold Measure | Response Category Name |
|---|---|---|---|---|---|---|---|
| | | | | Average measure | Outfit MnSq | | |
| ***Overall Impression*** | 1 | 0 | | | | | A1 |
| | 2 | 0 | | | | | A2 |
| | 3 | 44 | 22 | -2.69 | 1.0 | | B1 |
| | 4 | 90 | 45 | -1.65 | 0.7 | -2.86 | B2 |
| | 5 | 48 | 24 | 1.03 | 0.6 | 0.21 | C1 |
| | 6 | 16 | 8 | 2.34 | 0.8 | 2.65 | C2 |
| ***Range*** | 1 | 0 | | | | | A1 |
| | 2 | 0 | | | | | A2 |
| | 3 | 52 | 26 | -2.91 | 1.0 | | B1 |
| | 4 | 91 | 46 | -1.67 | 0.7 | -2.86 | B2 |
| | 5 | 41 | 21 | 1.04 | 0.7 | 0.36 | C1 |
| | 6 | 14 | 7 | 1.90 | 1.0 | 2.50 | C2 |
| ***Accuracy*** | 1 | 0 | | | | | A1 |
| | 2 | 0 | | | | | A2 |
| | 3 | 59 | 30 | -2.73 | 1.4 | | B1 |
| | 4 | 81 | 41 | -1.89 | 1.2 | -2.82 | B2 |
| | 5 | 49 | 25 | 0.33 | 1.7 | -0.14 | C1 |
| | 6 | 9 | 5 | 1.69 | 1.1 | 2.97 | C2 |
| ***Fluency*** | 1 | 0 | | | | | A1 |
| | 2 | 0 | | | | | A2 |
| | 3 | 77 | 39 | -2.77 | 1.2 | | B1 |
| | 4 | 65 | 33 | -1.72 | 0.8 | -2.09 | B2 |
| | 5 | 39 | 20 | 0.85 | 0.9 | 0.05 | C1 |
| | 6 | 17 | 9 | 1.57 | 1.2 | 2.05 | C2 |
| ***Coherence*** | 1 | 0 | | | | | A1 |
| | 2 | 0 | | | | | A2 |
| | 3 | 50 | 25 | -2.47 | 1.2 | | B1 |
| | 4 | 90 | 45 | -1.66 | 0.9 | -2.71 | B2 |
| | 5 | 39 | 20 | 1.19 | 0.7 | 0.49 | C1 |
| | 6 | 19 | 10 | 1.86 | 1.4 | 2.22 | C2 |

## ■5.0 DISCUSSION

To answer the first research question, findings show that the CEFR rating scales were properly functioned in four out of six categories which were from B1 to C2 (intermediate to advanced levels). Despite these learners being novice raters and first timers in being exposed to SAPA practice, they were able to use the rating scale usefully after a period of rater training. The measures reported (average measure, outfit mean-squares and Rasch-Andrich threshold) suggest that the participants did apply what they had learnt during training to the actual practice. Absence of A1 and A2 (beginner level) ratings was anticipated since the participants' level of English was modest. Although some may argue that the rating scale was flawed due to this absence, the researchers believed that it represents the users' ability in applying the rating scale. As mentioned in the review earlier, the CEFR should act as impetus in further development of rating scale to suit the context and requirements of a test or assessment (Davis, 2015). In the context of this study, the participants were immersed in A*a*L whereby they were the key assessor of their own performance. No teachers (or researchers) intervened during SAPA practice to influence the process in awarding the ratings. From the results, it suggests that for these participants, four level rating scale should suffice to assess their oral proficiency. In addition, this learner-oriented rating scale functioning also exemplifies the theory of measurement as argued by North.

To answer the second research question, the researchers found that participants were able to discriminate up to four levels of ratings (B1, B2, C1 and C2) for each criterion. However, the distribution of these ratings was uneven and generally, participants awarded B2 for four CEFR criteria (overall impression, range, accuracy and coherence) while for fluency, participants were inclined to award B1. This may imply that participants had certain biasness or severity in assessing fluency. This is also consistent with many studies which found that fluency was more observable in assessing speaking (Bosker *et. al*., 2012)as listeners were able to detect pauses or false starts easily.

## ■6.0 CONCLUSION AND IMPLICATIONS

In conclusion, only four (B1, B2, C1 and C2) out of six categories of the CEFR rating scale functioned usefully during SAPA practice by modest ESL learners on their oral proficiency. Consequently, scaling structure for each criterion also displayed similar pattern when participants only utilised intermediate and advanced levels of the rating scale. However, B2 seemed to dominate four criteria while for fluency, participants had the tendency to award B1. Based on these findings, it implies that modest ESL learners in this study were able to rate their performance by using the CEFR rating scale purposefully. However, the researchers would like to caution that not all modest ESL learners may be able to do so since the term modest is quite subjective as it depends on the assessment scale they were subjected on. Perhaps

future researchers could extend this study by involving beginner learners but another caution is warranted: descriptors may have to be modified or simplified to accommodate their limited understanding. In addition, since this study was conducted in A*a*L context, the results obtained indicate that modest ESL learners were capable of being the key assessor of their own oral proficiency. Although some may question the reliability and validity of learners' ratings, the researchers would like to reiterate that this practice is operationalised as a learning tool to empower learners in taking charge of their own learning. This practice is by no means a substitute for teacher assessment. Eventually, when learners learn to assess and use rating scales accordingly, it cultivates accountability for their own progress in oral proficiency.

## References


Alderson, J. C., Figueras, N., & Kuijper, H. (2006). Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference : The Experience of The Dutch CEFR Construct Project. *Language Assessment Quarterly*, *3*(1), 3–30.

Bosker, H. R., Pinget, A.F., Quene, H., Sanders, T., & de Jong, N. H. (2012). What makes Speech Sound Fluent? The Contributions Of Pauses, Speed And Repairs. *Language Testing*, *30*(2), 159–175.

Brantmeier, C., Vanderplank, R., & Strube, M. (2012). What About Me? *System*, *40*(1), 144–160.

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

Davis, L. (2015). The Influence Of Training And Experience On Rater Performance In Scoring Spoken Language. *Language Testing*. *33(1), 117-135*

Deygers, B., & Van Gorp, K. (2015). Determining the Scoring Validity Of A Co-Constructed CEFR-Based Rating Scale. *Language Testing*, *32*(4), 1–21.

Earl, L.M. (2003).*Assessment as Learning*. Thousand Oaks: Corwin Press.

Eckes, T. (2009). Many-Facet Rasch Measurement. In S. Takala (Ed.), *Reference Supplement To The Manual For Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, Assessment*, 2–54. Strasbourg, France: Council of Europe/Language Policy Division.

Falchikov, N. (1986). Product Comparisons and Process Benefits of Collaborative Peer and Self-Assessments. *Assessment & Evaluation in Higher Education.*11, 146–166.

Figueras, N. (2012). The Impact of the CEFR. *ELT Journal*, *66*(4), 477–485.

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.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective Rating Scale Development For Speaking Tests: Performance Decision Trees. *Language Testing*, *28*(1), 5–29.

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

Hand, D. (1996). Statistics and the Theory of Measurement. *The Royal Statistical Society*, *159*(3), 445–492.

Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & Hirvela, T. (2014). Assessing Learners' Writing Skills In A SLA Study: Validating The Rating Process Across Tasks, Scales And Languages. *Language Testing*, *31*(3), 307–328.

Hulstijn, J. H. (2007). The Shaky Ground beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency. *The Modern Language Journal*, *91*(4), 663–667.

Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P., & Florijn, A. (2011). Linguistic Competences Of Learners Of Dutch As A Second Language At The B1 And B2 Levels Of Speaking Proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*(2), 203–221.

Ibberson, H. (2012). *An Investigation Of Non-Native Learners Â€™ Self-Assessment Of The Speaking Skill And Their Attitude Towards Self-Assessment*. (Doctoral dissertation, University of Essex, UK). Retrieved from British Library E-Theses Online Service

Idris, M., & Zakaria, M. H. (2016). Gauging ESL Learners ' CEFR Ratings on Oral Proficiency in Rater Training. *Man in India*, *96*(6), 1675–1682.

Jones, N., & Saville, N. (2009). European Language Policy: Assessment, Learning, and the CEFR. *Annual Review of Applied Linguistics*, *29*(2009), 51–63.

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.

Linacre, J.M. (2014). Winsteps Rasch Measurement Version 3.71 [Software]. Retrieved from: http://www.winsteps.com.

Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching*, *39*(3), 167.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy. *The Modern Language Journal*, *91*(4), 645–655.

Lodico, M., Spaulding, D. &Voegtle, K. (2006). *Methods in Educational Research: From Theory to Practice*. San Francisco, CA: Jossey-Bass

North, B. (2014). *The CEFR in Practice. English Profile Studies, 4*. Cambridge: Cambridge University Press.

Nulty, D. (2011). Peer and Self-Assessment In The First Year Of University. *Assessment & Evaluation in Higher Education*, *36*(5), 493–507.

Remler, D. K. & Van Ryzin, G. G. (2015). *Research Methods in. Practice: Strategies for Description and Causation.* Thousand Oaks, CA: Sage.

Schoenfeld, A. H. (2016). 100 Years of Curriculum History, Theory, and Research. *Educational Researcher*, *45*(2), 105–111.

Topping, K. (1998). Peer Assessment Between Students in Colleges and Universities. *Review of Educational Research*, *68*(3), 249–276.

Tyler, R. W. (1949). *Basic Principles of Curriculum and Instruction*. Chicago: University of Chicago Press.

Valax, P. (2011). *The Common European Framework of Reference for Languages: A Critical Analysis Of Its Impact On A Sample Of Teachers And Curricula Within And Beyond Europe (*Doctoral dissertation, University of Waikato, New Zealand). Retrieved from http://researchcommons.waikato.ac.nz/handle/10289/5546