International Journal of Research in English Education (IJREE)

Original Article **Published online: 20 September 2022.**

Development and Validation of a Training-Embedded Speaking Assessment Rating Scale: A Multifaceted Rasch Analysis in Speaking Assessment

Houman Bijani[1]*, Bahareh Hashempour[2], & Salim Said Bani Orabah[3]

* Correspondence:
houman.bijani@gmail.com
1. Islamic Azad University, Zanjan Branch, Zanjan, Iran
2. University of Zanjan, Zanjan, Iran
3. University of Technology and Applied Sciences, Ibra, Oman

Received: 24 March 2022
Revision: 18 June 2022
Accepted: 11 July 2022
Published online: 20 September 2022

Abstract

Performance testing including the use of rating scales has become widespread in the evaluation of second/foreign oral language assessment. However, no study has used Multifaceted Rasch Measurement (MFRM) including the facets of test takers' ability, raters' severity, group expertise, and scale category, in one study. 20 EFL teachers scored the speaking performance of 200 test-takers prior and subsequent to a rater training program using an analytic rating scale consisting of fluency, grammar, vocabulary, intelligibility, cohesion, and comprehension categories. The outcome demonstrated that the categories were at different levels of difficulty even after the training program. However, this outcome by no means indicated the uselessness of the training program since data analysis reflected the constructive influence of training in providing enough consistency in raters' rating of each category of the rating scale at the post-training phase. Such an outcome indicated that raters could discriminate the various categories of the rating scale. The outcomes also indicated that MFRM can result in enhancement in rater training and functionality validation of the rating scale descriptors. The training helped raters use the descriptors of the rating scale more efficiently of its various band descriptors resulting in a reduced halo effect. The findings conveyed that stakeholders had better establish training programs to assist raters in better use of the rating scale categories of various levels of difficulty in an appropriate way. Further research could be done to make a comparative analysis between the outcome of this study and the one using a holistic rating scale in oral assessment.

Keywords: bias, interrater consistency, intrarater consistency, multifaceted rasch measurement (MFRM), rater training, rating scale

[DOR: 20.1001.1.25384015.2022.7.3.2.5]
[DOI: 10.52547/ijree.7.3.32]

[DOR: 20.1001.1.25384015.2022.7.3.2.5]

[DOI: 10.52547/ijree.7.3.32]

1. Introduction

During the last three decades, the application of language rating scales in the process of language proficiency assessment has become very popular. Moreover, they are now widely used to assess an individual learner's level of mastery over a particular skill and report the outcome (Huang, Huang, & Hong, 2016). When discussing rating process, we deal with the approach through which the rating system, most typically the rating scale is created and utilized. Language rating scales are now widely used to assess an individual learner's level of mastery over a particular skill and report the outcome (Trace, Janssen, & Meier, 2017). The application of rating scales needs interpretation by raters, and in case a number of raters are included, the reliability of the assigned scores may be influenced (McNamara, 1996). Variability amongst raters on account of the application of the scoring rubric is handled by rater training. However, this requires that the rating scale be well-constructed in advance so that it can discriminate test-takers consistently.

Ratings are done by studying and observing a sample of language performance under a testing condition. In particular, since speaking assessment is done by raters who are commonly trained, the use of rating scales seems essential. According to Flake (2021), the use of rating scales is crucial because of evidence from several pieces of research (e.g., Barkaoui, 2011; Bijani & Fahim, 2011; Huang, Bailey, Sass, & Shawn Chang, 2020; Theobold, 2021) demonstrated that even experienced raters may disagree first, about the nature of language sub-skills involved in the assessment of language ability, second, about which items measure what skills, and third, about the difficulty level of test items and tasks. Knoch (2009) identified the features of a good rating scale as 1. to be able to discriminate between various levels of performance assessment, 2. to be practical in the process of rating, and 3. practicality in a majority of performance test samples. As McNamara (1996) notes, a rating scale shows what skills, techniques, and capabilities are being evaluated by a test. For this cause, the enhancement of a rating scale and the descriptors for each scale level is important for the validity of measurement.

2. Literature Review

2.1 Rating Scales Used on Speaking Assessment

As Bachman (2004) argues, the scale which is used for evaluating performance tasks, like speaking tests, demonstrates the theoretical foundations based on which the test is established. A rating scale shows what skills, techniques, and capabilities are getting evaluated by a test. For such reason, the creation of a rating scale and the descriptors used for each level of the scale is important for the validity of measurement. In literature, two major sorts of rating scales are 1. *Analytic scales*, and *2. Holistic scales* (Luoma, 2004).

2.1.1 Holistic Rating Scale

Many assessment programs apply a holistic rating or the assigning of only one score to a performance based on the total perception of the performance (e.g., Attali, 2016; Kim, 2011). Holistic scoring has been widely used in oral assessment over the past 25 years and has a number of positive features. Considering a practical viewpoint, it is much faster and thus less costly to listen to an oral performance once and assign a single score. As Tavakoli, Nakatsuhara, and Hunter (2020) note, there are also other advantages to holistic scoring. They believe that holistic scoring is intended to focus on raters' attention on the strength of the performance, not on the deficiencies, so test-takers are rewarded for what they do well. They also argue that holistic rating is much more valid than analytic rating methods because in analytic scoring methods there is too much attention which distorts the meaning of the whole.

On the other hand, holistic scoring has several disadvantages, particularly in second language contexts. One disadvantage, according to Luoma (2004), is that holistic scoring does not provide efficient distinguishing information test takers' speaking proficiency since a single score does not permit raters to identify the difference between different facets of speaking like syntax, lexicon, organization, etc. One other drawback of holistic rating is that the interpretation of holistic scores is not always simple since raters reach the same score through different criteria. For instance, a certain performance might be given a 5 on a holistic scale by one rater because of its fluency, while another rater might give the same performance a 5 because of its linguistic features (Weigle, 2002).

2.1.2 Analytic Rating Scale

In analytic scoring, performances are scored on the basis of several facets of speaking criteria. Considering the intention of the evaluation, performances could be scored based on features like cohesion, fluency, grammar, lexicon, etc. The analytic rating provides more deep-down and detailed information about the test takers' performances in

various facets of speaking and for this reason, it is preferred over holistic scoring. One of the positive points of analytic rating over holistic one is that it provides more beneficial diagnostic data concerning students' speaking abilities (Winke & Gass, 2013). Analytic rating is more beneficial in rater training since experienced raters are better able to realize and apply the intended rating criteria than a holistic rating scale (Weigle, 2002). Analytic rating is especially fruitful for L2 learners. For example, performance may be quite well developed but have numerous grammatical errors. Knoch (2009) in an analytical study of rating scales distinguished six determining factors using the Diagnostic English Language Needs Assessment (DELNA) scoring rubric reflecting that analytic rating scales can classify test-takers based on more detailed strength and weakness levels. Bijani and Fahim (2011) in another analytical study on second language writing assessment realized that the use of an analytic rating scale could better distinguish test-takers based on various scoring criteria.

2.2 The Impact of a Rating Scale on Speaking Assessment

There has been more research about the impact of rating scales in speaking assessment. May (2009) compared the holistic and analytic evaluations of college students and professional speakers and found that professional speakers were distinguished from the college students on the analytic scale but not on the holistic scale. Regarding the relative reliability of different scale types, Davis (2016) found that analytic scores were more reliable as compared to holistic scores, although there was no rater training involved in either of these studies. In a research study, Khabbazbashi (2017) showed that 14 untrained raters scoring four audio cassettes of learners of Spanish differed basically in how they interpreted their perceptions of the points on ACTFL scale. She concluded that, without training, low inter-rater reliability was only what to be expected. After doing this, the raters changed their rating pattern in certain ways, particularly when using the rating scale which could escalate the inter-rater consistency. Developments or variations in the use of the rating scale were monitored through periodic rerating of a small number of oral tests. She further recommended the provision of training and retraining.

Attali (2016) investigated the rating of 40 essays written by Japanese students by employing 40 native English speakers. Each rater scored all the 40 essays on a six-point analytic scoring rubric of five categories. The results showed that some raters scored higher ability test-takers more severely and lower ability ones more leniently than what was predicted. Such a result indicated that various groups of raters used the rating scale differently. Kuiken and Vedder (2014) studied a group of experienced and inexperienced language teachers and provided both with one-day training in rating oral performance tests. Using a MFRM, they found that inexperienced raters' ratings were relatively severer than the experienced ones' concerning politeness and pronunciation and that they overfitted the given model, i.e., there was insufficient variability in their ratings. However, experienced raters were likely to have more diversity in their ratings, and severity when employing the scoring rubric. Kuiken and Vedder then concluded that there exist factors, other than the ones in the rating scale, which affect the raters' scoring. Nakatsuhara (2011) in a study of spoken language found that some raters were inconsistent in the particular categories of the rating scale. Some rated grammar more severely and some leniently. Some were severe towards vocabulary and fluency, whereas some others to task types.

Rater training programs can assist raters in better realizing the criteria and features of the rating scales which might influence heir rating behavior (Kim, 2015). Winke, Gass, and Myford (2012) found training effective in reducing rater variability. Accordingly, in the absence of rater training programs, raters with various levels of expertise may assign different scores to the language being tested (Bijani, 2010; In'nami & Koizumi, 2016). Thus, extended training programs will aid them to develop a common reference framework. Internal consistency, which is the goal of rater training programs, is tightly linked to the application of a specific scoring rubric (Sawaki, 2007). Since self-consistency, according to Khabbazbashi (2017), often cannot be obtained by rater training, it is assumed that what is significant in obtaining consistency is how successfully a rater achieves mastery over the guiding manuals of a specific scoring rubric. Similarly, Linacre (1989) used the term rater severity and variation amongst raters based on the way they interpret scoring rubrics.

The multi-faceted Rasch model (MFRM), introduced by Linacre (1989) which can be conducted using the computer software FACETS, takes a various approach to the concept of rater variation by not only exploring rater factors in performance-based language assessment but also by providing feedback to the raters on their scoring performance (Eckes, 2015). Bias analysis which can be applied via the MFRM can assist researchers to investigate and distinguish the origins of bias of raters; therefore, resulting in the development of raters' training and rating scale improvement (Fan & Yan, 2020). In this approach, rater variation is viewed as an unavoidable section of the scoring process, and

[DOR: 20.1001.1.25384015.2022.7.3.2.5]
[DOI: 10.52547/ijree.7.3.32]

instead of impeding assessment, is regarded as literally advantageous since it provides sufficient change to allow the possible estimation of rater severity, task difficulty, and test-taker ability through the use of a single linear scale.

Nevertheless, a majority of the studies done until now have explored the use of FACETS on just a couple of facets; for instance, the analysis of rater's severity/leniency on special test-takers (Huang, Huang, & Hong, 2016), and task types (Kyle, Crossley, & McNamara, 2016). On the contrary, no research, until now, has consisted of the facets of test takers' ability, raters' severity, group expertise, and scale criterion category all in just one study together with their impacts bilaterally. Although little research has investigated the variations between trained raters and untrained ones in oral assessment (e.g., Bijani, 2010; Gan, 2010; Kim, 2011), little research has employed a pre- and post-training research design exploring the effect of training on the reduction of raters' biases to the rating scale categories leading to increase in their consistency measures. Moreover, there have been very few studies investigating the impact of training in second language oral assessment since raters might change the way they interpret the categories of the scale (e.g., Bijani, 2010; Gan, 2010; Kim, 2011). Besides, levels of interrater agreement are still under question and raters may change in terms of consistency over each other. On the other hand, whether there is a decrease after training of individual biases related to scoring the rating scale categories and their difficulty levels is not clear.

Therefore, this study investigated the impact of the rater training program on their severity/leniency measures, consistency, and biases towards each rating scale category in a pre-, and post-training research design. This study is intended to account for the above-mentioned four facets by exploring the impact of the training program in reducing raters' biases in rating the categories of the rating scale for experienced and inexperienced raters. Moreover, this study investigated the criteria that raters use to judge the quality of learners' speaking ability with respect to the application of a particular rating scale, their interpretation of the use of the rubric categories, and the effects of training on the rating criteria and interpretation of the rubric. Consequently, the research question given below can be formulated:

RQ: Are the scale categories used in the study at the same level of difficulty?

3. Methodology

3.1 Design of the Study

In order to investigate the research question of this research, the researchers employed a pre-post, mixed-methods research design in which a combination of quantitative and qualitative approaches was used to investigate the raters' development over time concerning rating L2 speaking performance (Cohen, Manion, & Morrison, 2007). This method offered a comprehensive approach to the investigation of the research questions involving a comparison of raters' and test takers' perceptions before and after the rater training program. In addition, the type of sampling which was used in this study was "subjects of convenience", that is the subjects were selected based on certain reasons and they were not selected randomly (Dörnyei, 2007).

3.2 Participants

As many as 200 adult Iranian EFL students, consisting of 100 males and 100 females, between 17 and 44 years of age took part as test-takers. The students used in this research were chosen from the Iran Language Institute (ILI) studying at the intermediate, upper-intermediate, and advanced levels. 20 Iranian EFL teachers, consisting of 10 males and 10 females, between 24 and 58 years of age took part as raters. To do this research, the raters had to be separated into two groups of experienced raters and inexperienced ones to explore to what extent they are similar to and different from each other and whether one group would be superior to the other one or not. Consequently, a background questionnaire, adapted from McNamara and Lumley (1997), which was used to elicit the following information including (1) *demographic information*, (2) *rating experience*, (3) *teaching experience*, (4) *rater training*, and (5) *relevant courses passed* was awarded to the raters. These factors are presented in Table 1.

[DOR: 20.1001.1.25384015.2022.7.3.2.5]
[DOI: 10.52547/ijree.7.3.32]

Table 1. Criteria for rating expertise

Rater group	Criteria			
	Rating experience	Teaching experience	Rater training	Relevant courses passed
Inexperienced	Fewer than 2 years	Fewer than 5 years	Fewer than 2 years	Less than the four core courses • Pedagogical English grammar • Phonology and phonetics • SLA • Second language assessment
Experienced	Over 2 years with the use of both analytic and holistic scale	Over 5 years of teaching in different settings (e.g., various students age groups and various proficiency levels)	Over 2 years	All four main courses • Pedagogical English grammar • Phonology and phonetics • SLA • Second language assessment plus at least 2 courses of the selective courses.

Therefore, the raters were classified into two classes of expertise with respect to their experiences specified below.

A. Raters with no or fewer than two years of experience in rating and undertaking rater training, plus no or fewer than five years of experience in English language instruction and managed to pass fewer than the four main courses relevant to English language teaching. From now on these raters are referred to as NEW raters.

B. Raters with two and more years of rating experience and undertaking rater training, plus five and more years of experience in English language teaching and managed to pass the whole four main courses relevant to English language teaching as well as a minimum of two other selective courses. From now on these raters are referred to as OLD raters.

3.3 Instruments

3.3.1 Oral Tasks

The oral proficiency of test-takers was assessed via five various tasks consisting of description, narration, summarizing, role-play, and exposition tasks. Task 1 (*Description Task*) displays test takers' background information and their individual experience in giving responses when they are not provided with any input. Moreover, tasks 3 (*Summarizing Task*) and 4 (*Role-play Task*) display the listening ability of test-takers in oral responses. Tasks 2 (*Narration Task*) and 5 (*Exposition Task*) require the test takers to give responses to pictorial prompts consisting of a series of pictures, figures, graphs, and tables.

3.3.2 Scoring Rubric

The task performance of each test taker was evaluated through Educational Testing System (ETS, 2001) analytic scoring rubric. In the scoring rubric of ETS (2001), each task is evaluated through the use of the following criteria consisting of *fluency*, *grammar*, *vocabulary*, *intelligibility*, *cohesion*, and *comprehension*. Each of these criteria is accompanied by a set of 7 descriptors.

3.4 Procedure

3.4.1 Pre-training Phase

Before gathering any data from the test-takers, the raters' background questionnaire was awarded to them to complete. This was done in order to assist the researchers to divide the raters into experienced and inexperienced ones. In order

[DOR: 20.1001.1.25384015.2022.7.3.2.5]
[DOI: 10.52547/ijree.7.3.32]

to run the speaking tasks, the 200 test takers were classified randomly into experienced and inexperienced ones in a way that half of them participated in each phase of the research, i.e., pre- and post-training phases.

3.4.2 Rater Training

Following the pre-training phase of the study, the raters took part in a rater training session during which they were familiarized with the speaking tasks and the scoring rubric. Also, they were given time to exercise the instructed materials. On top of that, the raters each were provided with feedback with regard to their previous scorings. Accordingly, the raters with z-scores beyond ±2 were regarded as significantly biased ones thus they were reminded of their bias individually. In terms of the consistency of the raters, infit mean square has an acceptable range which is between 0.6 and 1.4 (Wright & Linacre, 1994); thus, any measure beyond this range is regarded as misfitting in such a way that infit mean square values lower than 0.6 are regarded as too consistent (overfit the standard model) and the ones over 1.4 as inconsistent (underfit the standard model).

3.4.3 Post-training Phase

Right after the rater training program, the speaking tasks were once again run. As was mentioned before in the pre-training data collection procedure, the remaining second half of the test takers (consisting of 100 individuals) were employed for data elicitation.

3.5 Data Analysis

To analyze the findings of the research question, a pre-post method design was adopted to explore the raters' progress in scoring second language oral performance (Cohen, Manion, & Morrison, 2007). Having collected quantitative data, they were analyzed with MFRM within two rating sessions for the four test facets including test-takers, rater and rater group, and scoring criterion as well as the interactions amongst them to identify any possible variation in the behavior of the raters and their biasedness. The pattern of scoring of the two groups of raters (inexperienced & experienced) was investigated as they rated test takers' oral productions. The quantitative data were analyzed (1) across the two rater groups to explore their capability in a cross-sectional pattern at each scoring node, and (2) within each rater group to investigate the progress of the raters' capability. The interactional effect of the raters of both groups of expertise with rating scale categories was investigated to identify any hypothetical differences with respect to the impact of training between the two groups in their assessment of rating scale categories.

4. Results

Table 2 demonstrates the average grades awarded by the raters of the two groups of expertise to the performance of test-takers in each category of the rating scale prior to the training program. The table depicts that NEW raters were more lenient than OLD ones and thus awarded higher grades than OLD raters.

Table 2. Descriptive statistics of grades awarded by raters to test takers' oral performance on each rating scale category (Pre-training)

Tasks	N	Mean			SD (Both)
		NEW	OLD	Both	
Cohesion	100	3.64	3.03	3.33	0.36
Intelligibility	100	3.96	3.28	3.62	0.24
Fluency	100	4.41	3.78	4.09	0.27
Comprehension	100	4.76	4.17	4.46	0.08
Vocabulary	100	6.08	5.46	5.77	0.11
Grammar	100	6.12	5.57	5.84	0.22
Mean		4.82	4.21	4.51	0.21
SD		1.05	1.08	1.06	0.10

[DOR: 20.1001.1.25384015.2022.7.3.2.5]
[DOI: 10.52547/ijree.7.3.32]

Moreover, to understand whether there exists a significant difference in raters' rating of the oral productions of test-takers, a one-way ANOVA was run on the scale categories of the rating scale (Wright & Linacre, 1994). Table 3 displays the outcome of one-way ANOVA related to the raters' rating of test takers' oral performances on each scale category.

Table 3. One-way ANOVA of raters' rating of test takers' oral performance capability on each rating scale category (Pre-training)

	Sum of Squares	*df*	Mean Square	F	Sig.
Between Groups	166.66	5	33.33	125.05	0.000
Within Groups	158.33	594	0.267		
Total	325.00	599			

$p<0.05$

The result of the above table demonstrates that there exists a significant mean difference in terms of raters' rating of the oral production of test-takers on each category of the rating scale at the pre-training phase of the study. The outcomes implied that there existed a significant mean difference between all pairs of the rating scale categories in terms of their ratings of test takers' oral performance capability at the pre-training phase with the exception of these pairs: vocabulary-grammar (p = 0.890), and cohesion-intelligibility (p = 0.222).

To evaluate the validity measure of the analytic descriptors, MFRM was used. MFRM is beneficial in the validation of a scoring rubric so that it is possible to analyze the sources of scoring variation. Besides, bias analysis examines the systematic sub-pattern interaction between raters and the scoring rubric (Schaefer, 2008). A bias analysis was administered for the difficulty measurement of the categories in rating to observe whether particular raters treat any of the rating scale categories with bias, i.e. rating them severely or leniently. FACETS software is proficient enough to measure biases for every one of the categories of the scoring rubric via juxtaposing the expected and observed scores in a dataset and afterwards notifying the result as residuals. Residuals are then converted into z-scores which display the bias value. Such z-score displays the amount of significant deviation expected from that specific rater for an acceptable and routine score change. A z-score between the range of -2 and +2 is considered an admissible range of bias in terms of a rater's behavior (McNamara, 1996). Table 4 represents the difficulty measurement report for the rating scale scoring categories at the pre-training stage.

Column one (Scale category) displays the rating scale employed in this study. *Column two (scale difficulty)* pictures the difficulty of scale categories from cohesion, as the most severely rated scale category, (difficulty logit: 0.79) to grammar, as the least severely rated scale category (difficulty logit: -0.46). The *third column (SE)* shows the standard error which is small here (from 0.03 to 0.04 logits). This indicates the high precision of measurement.

The *fourth column (Infit MnSq.)* is also regarded as *"quality control fit statistics"* that reflects the degree to which the data suit the Rasch model. The expected value of fit statistics is 1 and ranges from zero to infinity. Nevertheless, there is no precise range for the interpretation of fit statistics; thus, the acceptable range of fit is done judgmentally not statistically. According to Myford and Wolfe (2004), such resolutions are highly dependent on the assessment context. Wright and Linacre (1994) suggest a standard range between 0.6 and 1.4 logit values. Thus, to examine the fit statistics of the raters, it was employed. Fit values within the acceptable range indicate that no category was indicated as misfitting, thus any value beyond the acceptable limit displays misfitting. The categories lower than this range are *overfit* or *too consistent*, indicating that they were rated too consistent which indicates that the raters had difficulty separating the different scale categories, in other words, they do not get the use of the whole range of the scale category. In the same vein, the values higher than this range are *underfit* (misfit) or *too inconsistent*, indicating that they were rated too inconsistently. In literature, the terms underfit and overfit are often both referred to as misfit (Eckes, 2015). Here, cohesion (Infit MnSq = 1.5) was specified as misfitting. This reflects that this category was scored too inconsistently by the raters before training.

[DOR: 20.1001.1.25384015.2022.7.3.2.5]
[DOI: 10.52547/ijree.7.3.32]

Nevertheless, the logit difficulty measures do not alone indicate whether the variation is consequential; consequently, FACETS gives other criteria of reliability variation for each facet. ***Separation index***, ***Reliability***, and ***Fixed chi-square*** are amongst the most informative ones given accordingly. The ***separation index*** is used to determine the spread values of the measures relevant to the exactness of which. Provided that the scale categories were at the same level of difficulty, the standard deviation of the scale category difficulty measures had better be equal to or smaller than 1.00. Here, cohesion was diagnosed as the most severely rated category (difficulty logit: 0.79), and grammar, as the least severely rated category (difficulty logit: -0.46), hence resulting in the separation index of 1.25. The ***reliability*** for the scoring rubric categories displays the proportion of consistency amongst the raters in scale category difficulty. The high value of *the reliability index* (r = 0.90) demonstrates that the analysis managed to reliably break up the categories of the rating scale into diverse difficulty levels. ***Fixed chi-square*** scrutinizes the extent to which the facet components are equally rated or not. The fixed chi-square measure for the entire six categories of the rating scale was calculated. The chi-square value specifies whether there existed a significant difference in the difficulty levels of the rating scale categories ($X^2_{(5, N=6)}$ = 7464.12, p<0.00). Here, a Chi-square with high values specifies that at least a minimum of two categories of the rating scale did not have an overlap on a component (e.g, difficulty). Conclusively, the result indicated that the categories of the scoring rubric did not share the same level of difficulty level.

Consequently, the *separation in the category difficulty* is rather high (1.25 logits) which shows that the categories of the scoring rubric were more than 1 statistically diverse degree of difficulty with high *reliability of the separation index* (0.90). The high reliability shows that the scale categories were reliably separated concerning their difficulty degree and that the analysis was reliable. As argued by Wink, Gass, and Myford (2012), the indices of separation reliability which are close to zero display that the scale categories did not vary substantially with regard to their measures of difficulty. Nevertheless, any value adjacent to1.0 depicts that the categories were reliably broken up concerning their difficulty measures. The *fixed chi-square value* for the whole six scale categories was measured ($X^2_{(5, N=6)}$ = 7464.12, p<0.00); therefore the null hypothesis, indicating that the scale categories were equally difficult, would be rejected. In other words, the finding shows that there is a significant variation among the six rating scale categories with respect to difficulty in the pre-training phase. This finding tells us that the raters consistently rated cohesion more severely than the other categories, whereas, in contrast, they rated grammar more leniently than the other categories. In other words, the raters tended to be harsher (less tolerant of weaknesses) on cohesion, intelligibility, fluency, and comprehension, whereas they tended to assign higher grades to test takers on vocabulary and grammar. The fact that the scoring rubric categories were rated having different difficulty measures show that the raters were not treating them similarly during the rating. This means that the raters could discriminate among them.

Table 4. Difficulty measurement report for the categories of the rating scale (Pre-training)

Scale category	Scale difficulty (logits)	SE	Infit MnSq.
Cohesion	0.79	0.03	1.5
Intelligibility	0.69	0.03	1.03
Fluency	0.44	0.03	0.7
Comprehension	0.17	0.04	0.6
Vocabulary	-0.41	0.03	1.1
Grammar	-0.46	0.03	0.9
Mean	0.20	0.03	0.97
SD	0.54	0.00	0.32
Fixed chi-square: 7464.12, *df*= 5, p<0.00			
Scale category separation index: 1.25			
Reliability index: 0.90			

[DOR: 20.1001.1.25384015.2022.7.3.2.5]
[DOI: 10.52547/ijree.7.3.32]

Table 5 demonstrates the average grades that the raters awarded to each group of expertise to the performance of test-takers of every one of the six rating scale categories employed at the post-training stage. The table, like before, displays that NEW raters were more lenient than OLD raters and thus awarded higher grades.

Table 5. Descriptive statistics of grades awarded by raters to test takers' oral production on each scale category (Post-training)

Tasks	N	Mean			SD (Both)
		NEW	OLD	Both	
Cohesion	100	4.27	4.02	4.14	0.13
Intelligibility	100	4.42	4.20	4.31	0.17
Fluency	100	4.85	4.51	4.68	0.13
Comprehension	100	5.06	4.68	4.87	0.12
Vocabulary	100	5.93	5.66	5.79	0.08
Grammar	100	5.71	5.57	5.64	0.10
Mean		5.04	4.77	4.90	0.12
SD		0.67	0.69	0.68	0.03

Additionally, a one-way ANOVA on the scale categories was carried on to identify the extent to which there exists a significant difference in raters' rating of the oral production by test takers following training. Table 6 displays the outcome of one-way ANOVA of the raters' grading of the oral production of test takers' on each category of the rating scale at the post-training phase of the study.

Table 6. One-way ANOVA of raters' rating of test takers' oral production on each scale category (Post-training)

	Sum of Squares	*df*	Mean Square	F	Sig.
Between Groups	238.427	5	47.685	367.978	0.000
Within Groups	76.975	594	0.130		
Total	315.402	599			

$p<0.05$

The result specifies that there exists a significant mean difference in terms of raters' grading of test takers' oral production on each scale category at the post-training phase. The outcome of the post hoc Scheffé test displayed that there existed a significant mean difference between all pairs except for fluency-comprehension (p=0.477). A bias analysis was administered for the difficulty measurement of the categories in rating. Table 7 represents the difficulty measurement report for the rating scale scoring categories in the post-training phase.

The *second column (scale category)* represents the scale categories from cohesion, as the most severely-scored category, (difficulty logit: 0.62) to vocabulary, as the least severely-scored category, (difficulty logit: -0.24). The *separation index* in the category difficulty was rather high (0.83 logits) and the *reliability of the separation index* was high (0.87). The *fixed chi-square* value for the entire six scale categories was calculated ($X^2_{(5, N=6)} = 811.63$, $p<0.00$); therefore, the null hypothesis, indicating that the scale categories were equally difficult, was rejected. In other words, the finding shows that there is a significant variation among the six rating scale categories regarding the difficulty in the post-training phase. Through making a comparison between the two phases of the study, the range of scale category

[DOR: 20.1001.1.25384015.2022.7.3.2.5]
[DOI: 10.52547/ijree.7.3.32]

difficulty measure was reduced considerably. This finding once again indicates that the raters consistently rated cohesion more severely than the other categories, whereas, in contrast, they rated vocabulary more leniently than the other categories. In other words, the raters were more likely to be harsher (less tolerant of weaknesses) on cohesion, intelligibility, fluency, and comprehension, whereas they tended to award higher grades to test takers on grammar and vocabulary.

The *fourth column (Infit MnSq.)* shows that, after training, no category was identified misfitting (beyond the tolerable range of 0.6 and 1.4 logit values). This reflects the constructive influence of the training program in providing enough consistency in raters' rating of each category of the rating scale in the post-training phase.

Table 7. Difficulty measurement report for the categories of the rating scale (Post-training)

Scale category	Difficulty (logits)	SE	Infit MnSq.
Cohesion	0.62	0.04	1.4
Intelligibility	0.47	0.02	0.7
Fluency	0.19	0.04	0.7
Comprehension	0.16	0.03	0.8
Vocabulary	-0.24	0.05	1.2
Grammar	-0.13	0.03	1.3
Mean	0.17	0.03	1.01
SD	0.33	0.01	0.31
Fixed chi-square: 811.63, *df*= 5, *p*<0.00			
Scale category separation index: 0.83			
Reliability index: 0.87			

Figure 1 plots graphically the information about scale category difficulty measure in the form of z-scores. The scale categories are placed on the horizontal axis and the z-scores on the vertical axis. It shows to what extent the scale categories were severely or leniently scored by the raters in the two phases of the study.

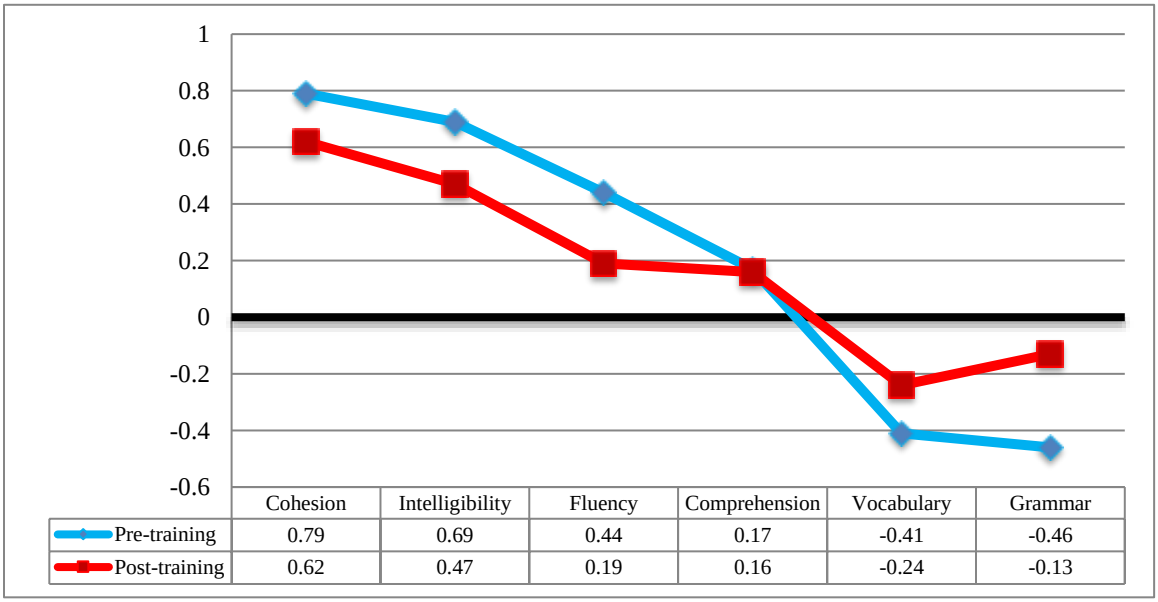

Figure 1. Scale category difficulty measure before and after training

[DOR: 20.1001.1.25384015.2022.7.3.2.5]
[DOI: 10.52547/ijree.7.3.32]

[DOR: 20.1001.1.25384015.2022.7.3.2.5]

[DOI: 10.52547/ijree.7.3.32]

5. Discussion

The research question dealt with whether the rating scale categories used in the study were at the same level of difficulty or not. The outcome of the study demonstrated that the categories were at different levels of difficulty even after the training program. However, this outcome by no means indicated the uselessness of the training program since the analysis of infit mean square values reflected the constructive influence of the training program in providing enough consistency in raters' rating of each category of the rating scale at the post-training phase. Such a result is parallel with the one conducted by Bijani and Fahim (2011) and Davis (2016) who found an increase in the consistency when raters used the scoring rubric categories after training. However, both these studies used ANOVA in data analysis.

In general, the differences concerning the difficulty of the scale categories and fit statistics indicated that raters managed to discriminate the various categories of the scoring rubric. The results displayed that the scoring rubric descriptors were not stable in a way that they changed radically concerning raters' scoring from before training to after training. This outcome supports the variable competence model of SLA by Tarone (1983) which indicates that an individual's understanding of language is variable. It must also be indicated that this point of view is also confirmed in language testing by Bachman (2004).

The outcomes, through considering the raters' infit mean square values indicated that the raters achieved higher levels of consistency after training. Likewise, raters' biases to the rating scale categories, as shown by z-scores, were reduced to a considerable extent which confirms the constructive impact of the training program in helping raters achieve higher measures of consistency as well. The fact that the raters had a great deal of leniency/severity in scoring using the rating scale categories, specifically in the pre-training phase, might most probably refer to their variability in interpreting the meaning of each criterion and its relevant descriptor. Still, a relatively similar outcome of rater training was observed showing raters' substantial differences which even training could not be effective enough to eliminate. The consequence of the research indicated that the raters improved consistency and reduced leniency/severity and bias after the training program concerning the use of the categories of the scoring rubric. The remaining differences regarding bias measures could probably be attributed to the halo effect which is due to the different ways of interpreting the scoring rubrics which is due to raters' confusion in the accurate application of the scale category descriptors or their overconcentration on a particular category. This outcome is fairly in line with that of (Bijani & Fahim, 2011; Khabbazbashi, 2017; Theobold, 2021; Winke, Gass, & Myford, 2012) who found similar outcomes in their research.

The outcomes of this study indicated that MFRM can be employed to investigate raters' rating behavior and can result in enhancement in rater training and validation of the functionality of the rating scale descriptors. This results in the use of a rating scale consistently by raters. MFRM data analysis lets us identify which categories of the scoring rubric could perfectly be set apart by the raters. That is, MFRM can be employed to assist test developers whether the creation of scale descriptors is employed the way they were intended (In'nami & Koizumi, 2016) or not. MFRM can point out sources of raters' bias thus making assessment fairer. It can diminish the intimidation of getting either confirmed or dismissed concerning the factors having nothing to do with their true ability. Besides, it can determine raters' bias by showing the extent to which raters show interaction with the categories of the rating scale. This can provide feedback to assist the raters to use rating scales more consistently. The result of fit statistics at the post-training stage implied that the raters managed to employ the scoring rubric descriptors consistently to score test takers' oral performances despite the various observed levels of severity.

With respect to the analytic scoring rubric benefited in the present research, the outcome of data analysis demonstrated that the rating scale descriptors provided raters with sufficient detailed information based on which to make decisions on test takers' oral proficiency when assigning scores. Some studies demonstrated that raters showed a halo effect when they faced problems using rating scales. For example, Fan and Yan (2020) argued that as raters cannot specify specific matters of rating scales, they resort to a more global and holistic use of them. This phenomenon was also observable in this study; however, the training program proved to be effective enough in reducing this halo effect providing raters with more explicit rating factors. That is, the training helped raters use the descriptors of the scoring rubric more efficiently of its various band descriptors because if raters use the scores centered around the middle of the scale, it will be less useful for test takers when they are presented with their performance profile. This outcome is parallel with that of Huang, Bailey, Sass, and Shawn Chang (2020) who found the constructive impact of training on the use of rating scale by raters. This finding shows that the halo effect does not certainly attribute to rater impact but

can be a result of the scoring rubric. Nevertheless, no matter what kind it is, the effect can be reduced, although not neutralized totally by rater training programs. A similar halo effect was observed by Kim (2015) in a study on rating test takers' oral ability.

6. Conclusion

The outcome of the study showed that training cannot easily eradicate raters' differences related to their characteristics. This is something that through more training and individual feedback could be better paved but not thoroughly removed. Scales have limited validity because they are unable to describe an oral performance adequately, therefore, the role of training is to clarify the vague points of a scale thus making its constructs valid enough for raters to use. On account of the rating scale descriptors analysis, the outcome of the study can not only inform teachers to concentrate on the areas in which students are weak but also can focus raters' attention on particular components of the rating scale to improve interrater reliability. It might be the case that the relatively high obtained reliability is due to the scoring rubric used in the study. Perhaps employing a valid scoring rubric has benefitted raters achieve consistency in scoring. Thus, it is suggested that similar research be conducted with a holistic rating scale or even without a scoring rubric to observe the possible contribution of a scoring rubric in training.

The findings of this research have several practical implications for testing and assessment organizations and other related stakeholders. The analysis of the norming sessions showed that some of the criteria in the scoring rubric were rather vague for inexperienced raters to use, thus they should be instructed to the raters orally during the training program. On account of the rating scale descriptors analysis, the outcome of the study can not only inform teachers to concentrate on the areas in which students are weak, but also it can focus raters' attention on particular components of the rating scale to improve interrater reliability. This research focused on an analytic rating scale having six categories for assessing oral performance assessment. Further research could be done by comparing the use of a holistic and an analytic scoring rubric thus determining which scale assesses test takers' oral ability more reliably. Besides, future studies can focus on other types of analytic rating scales having different descriptors or even developed rating scales to see which analytic rating scale better distinguishes the test takers concerning their performance ability into diverse levels.

References

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99-115. https://doi.org/10.1177/0265532215582283

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study on their veridicality and reactivity. *Language Testing, 28*(1), 51-75. https://doi.org/10.1177/0265532210376379

Bijani, H. (2010). Raters' perception and expertise in evaluating second language compositions. *The Journal of Applied Linguistics, 3*(2), 69-89.

Bijani, H., & Fahim, M. (2011). The effects of rater training on raters' severity and bias analysis in second language writing. *Iranian Journal of Language Testing, 1*(1), 1-16.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. London: Routledge.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135. https://doi.org/10.1177/0265532215582282

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies.* Oxford: Oxford University Press.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement*. Frankfurt: Peter Lang Edition.

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: a narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology, 11*(1), 1-14. https://doi.org/10.3389/fpsyg.2020.00330

Flake, J. K. (2021). Strengthening the foundation of educational psychology by integrating construct validation into open science reform. *Educational Psychologist, 56*(2), 132-141. http://doi.org/10.1080/00461520.2021.1898962

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher-and lower-scoring students. *Language Testing*, *27*(4), 585-602. doi:10.1177/0265532210364049

Huang, B. H., Bailey, A. L., Sass, D. A., & Shawn Chang, Y. (2020). An investigation of the validity of a speaking assessment for adolescent English language learners. *Language Testing*, *37*(2), 1-28. https://doi.org/10.1177/0265532220925731

Huang, H., Huang, S., & Hong, H. (2016). Test-taker characteristics and integrated speaking test performance: A path-analytic study. *Language Assessment Quarterly*, *13*(4), 283-301. https://doi.org/10.1080/15434303.2016.1236111

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*, *33*(3), 341-366. https://doi.org/10.1177/0265532215587390

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing*, *34*(1), 23-48. https://doi.org/10.1177/0265532215595666

Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking assessment.* Unpublished PhD thesis, University of Columbia.

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, *12*(3), 239-261. https://doi.org/10.1080/15434303.2015.1049353

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275-304. https://doi.org/10.1177/0265532208101008

Kuiken, F., & Vedder, I. (2014). Raters' decisions, rating procedures, and rating scales. *Language Testing*, *31*(3), 279-284. https://doi.org/10.1177/0265532214526179

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*(3), 319-340. https://doi.org/10.1177/0265532215587391

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.

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

May, L. **(**2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*(3), 397-421. https://doi.org/10.1177/0265532209104668

McNamara, T. F. (1996). *Measuring second language performance.* London: Longman.

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing*, *14*(2), 140-156. https://doi.org/10.1177/026553229701400202

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement. *Journal of Applied Measurement*, *5*(2), 189-227.

Nakatsuhara, F. (2011). Effect of test-taker characteristics and the number of participants in group oral tests. *Language Testing*, *28*(4), 483-508. https://doi.org/10.1177/0265532211398110

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing*, *24*(3), 355-390. https://doi.org/10.1177/0265532207077205

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493. https://doi.org/10.1177/0265532208094273

Tarone, E. (1983). On the variability of interlanguage systems. *Applied Linguistics*, *4*(2), 142-164. https://doi.org/10.1093/APPLIN/4.2.142

Tavakoli, P., Nakatsuhara, F., & Hunter, A. M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal*, *104*(1), 169-191. https://doi.org/10.1111/modl.12620

Theobald, M. (2021). Self-regulated learning training programs enhance university students' academic performance, self-regulated learning strategies, and motivation: A meta-analysis. *Contemporary Educational Psychology, 66*, 101976. https://doi.org/10.1016/j.cedpsych.2021.101976

Trace, J., Janssen, G., & Meier, V. (2017). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing, 34*(1), 3-22. https://doi.org/10.1177/0265532215594830

Weigle, S. C. (2002). *Assessing Writing.* Cambridge: Cambridge University Press.

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly, 47*(4), 762-789. https://doi.org/10.1002/tesq.73

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231-252. https://doi.org/10.1177/0265532212456968

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 369-386.

[DOR: 20.1001.1.25384015.2022.7.3.2.5]

[DOI: 10.52547/ijree.7.3.32]