Iranian Journal of Applied Linguistics (IJAL)

Vol. 26, No. 2, September 2023, 48-80

Construct Validation of a Rating Scale through a Training Program: A Multifaceted Rasch Analysis in Speaking Assessment

Wander Lowie

Ph.D., Department of Applied Linguistics, Groningen University, Netherland

Houman Bijani *

Ph.D., Department of English, Zanjan Branch, Islamic Azad University, Zanjan, Iran

Mohammad Reza Oroji

Ph.D., Department of English, Zanjan Branch, Islamic Azad University, Zanjan, Iran

Zeinab Khalafi

Ph.D. Candidate, Department of English, Zanjan Branch, Islamic Azad University, Zanjan, Iran

Pouya Abbasi

Ph.D. Candidate, English Department, Chabahar Maritime University, Chabahar, Iran

Abstract

Performance testing including the use of rating scales has become highly widespread in the evaluation of second/foreign oral assessment. However, few studies have used a pre-, post-training design investigating the impact of a training program on the reduction of raters' biases to the rating scale categories resulting in increase in their consistency measures. Besides, no study has used MFRM including the facets of test takers' ability, raters' severity, task difficulty, group expertise, scale category, and test version all in a single study. 20 EFL teachers rated the oral performances produced by 200 test takers before and after a training program using an analytic rating scale including fluency, grammar, vocabulary, intelligibility, cohesion and comprehension categories. The outcome of the study indicated that MFRM can be used to investigate raters' scoring behavior and can result in enhancement in rater training and validating the functionality of the rating scale descriptors. Training can also result in higher levels of interrater consistency and reduced levels of severity/leniency; however, it cannot turn raters into duplicates of one another, but can make them more self-consistent. Training helped raters use the descriptors of the rating scale more efficiently of its various band descriptors resulting in reduced halo effect. Finally, the raters improved consistency and reduced rater-scale category biases after the training program. The remaining differences regarding bias measures could probably be attributed to the result of different ways of interpreting the scoring rubrics which is due to raters' confusion in the accurate application of the scale.

* *Corresponding author:* Department of English, Zanjan Branch, Islamic Azad University, Zanjan, Iran. Email address: houman.bijani@gmail.co

Keywords: Bias; Interrater consistency; Intrarater consistency; Multifaceted Rasch Measurement (MFRM); Rater training; Rating scale

1. Introduction

During the last three decades, the use of language rating scales in the process of language proficiency assessment has become very popular. When discussing scoring procedures, we are concerned with the way in which the scoring system, most commonly the rating scale, is developed and used. Language rating scales are now widely used to assess individual learner's level of mastery over a particular skill and report the outcome (Bridley, 1998). The use of rating scales requires interpretation by raters, and in case several raters are involved, the reliability of test taker scores can be influenced (McNamara, 1996). Variability among raters on account of the use of the rating scales 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.

According to Luoma (2004) the use of rating scales is crucial because evidence from a number of studies (e.g., Barkaoui, 2011; Bijani & Fahim, 2011; Chalhoub-Deville, 1995; Knoch, 2007; Sawaki, 2007) 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. being able to discriminate between various levels of performance assessment, 2. being practical in the rating process, and 3. practicality in most performance test samples. As McNamara (1996) notes, a rating scale shows what skills or abilities are being measured by a test. For this reason, the development of a scale and the descriptors for each scale level is important for the validity of measurement. In literature, two main types of rating scales are discussed: (1) *Holistic scales*, and (2) *Analytic scales* (Luoma, 2004).

In analytic scoring, performances are rated on several aspects of speaking criteria. Depending on the purpose of the assessment, performances might be rated on features such as cohesion, fluency, grammar, vocabulary, etc. Analytic scoring provides more detailed information about the test takers'

performances in different aspects of speaking and for this reason it is preferred over holistic scoring. One of the advantages of analytic scoring over holistic one is that it provides more useful diagnostic information about student's speaking abilities (Bazyar, 2023; Winke & Gass, 2013). Analytic scoring is more useful in rater training as experienced raters can more easily understand and apply the criteria than a holistic scale (Weigle, 2002). Knoch (2009) in an analytical study of rating scales distinguished 6 determining factors using the Diagnostic English Language Needs Assessment (DELNA) rating scale reflecting that analytic rating scales can classify test takers based on more detailed strength and weakness levels. 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, O'Sullivan, Weir and Saville (2002) found that analytic scores were more reliable than holistic scores, although there was no rater training involved in either of these studies.

Internal consistency, which is the target of rater training, is also closely related to the use of a particular rating scale (Sawaki, 2007). Since self-consistency, according to Lumley and McNamara (1995), often cannot be obtained by rater training, it is assumed that what is important in obtaining consistency is how well a rater masters the guidelines of a special rating scale. On the other hand, the Multi-faceted Rasch model (MFRM) introduced by Linacre (1989), which can be done using the computer software FACETS, takes a different approach to the phenomenon of rater variation by not only investigating rater factors in performance-based language assessment but also by providing feedback to the raters on their rating performance (Lumley & McNamara, 1995). In this approach, rater variation is seen as an inevitable part of the rater process, and rather than being an obstacle to measurement, is considered actually beneficial because it provides enough variability to allow probabilistic estimation of rater severity, task difficulty, and test taker ability using the same linear scale.

However, most of the studies conducted so far have investigated the application of FACETS on only one or two facets. For example, the study of rater's severity/leniency on specific test takers (Lynch & McNamara, 1998), and on task types (Fulcher, Davidson & Kamp, 2011). However, no study, so far, has included the facets of test takers' ability, raters' severity, task difficulty, group expertise, scale criterion category, and test version all in a single study along with their bilateral effects. While a few studies have looked at the differences between trained and untrained raters in speaking assessment (e.g., Bijani, 2010; Gan, 2010; Kim, 2011; Maldar, 2022), few studies have used a pre- and post-training design

investigating the impact of training on reduction of raters' biases to the rating scale categories resulting in increase in their consistency measures. Besides, there have been very few studies exploring the effectiveness of training in second language speaking assessment because raters might vary in the way they interpret the categories of the scale. Moreover, levels of interrater agreement are still under question and raters may vary in terms of consistency over each other. On the other hand, whether there is a reduction following training of individual biases in relation to the scoring categories of the rating scale and their difficulty levels is not clear.

Therefore, this study investigated the effect of the rater training program on their severity/leniency measures, consistency and biases towards each rating scale category in a pre-, post-training research design. This study aimed to account for the above-mentioned six facets exploring the impact of the training program in the reducing raters' biases in scoring rating scale categories 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 use of a particular scoring rubric, their interpretation of the use of the rubric categories, and the effects of training on the rating criteria and interpretation of the rubric. Therefore, the following research questions can be formed:

RQ1: Are the scale categories used in the study at the same level of difficulty?

RQ2: Is there a reduction of rater biases in relation to the scoring each rating scale category following training?

RQ3: How does each rater's interactional bias to each rating scale category change throughout the study?

RQ4: To what extent are the rating scale categories similar or dissimilar to each other in rating and how does training affect their interrelationship?

2. Methodology

2.1. Participants

Two hundred (200) adult Iranian students of English as a Foreign Language (EFL), including 100 males and 100 females, ranging in age from 17 to 44 participated as test takers. The students were selected from Intermediate, Upper-intermediate, and Advanced levels studying at the Iran Language Institute (ILI).

Twenty (20) Iranian EFL teachers, including 10 males and 10 females, ranging in age from 24 to 58 participated as raters. In order to fulfill the requirements of this study, the raters had to be classified into two groups of experienced and inexperienced raters to investigate the similarities and differences among them and the likelihood advantages of one group over the other one; therefore, a background questionnaire, adapted from McNamara and Lumley (1997), eliciting the following information including (1) *demographic information*, (2) *rating experience*, (3) *teaching experience*, (4) *rater training* and (5) *relevant courses passed* was given to the raters. Thus, raters were divided into two levels of expertise on the basis of their experiences outlined below.

A. Raters who had no or less than two years of experience in rating and receiving rater training, and had no or less than 5 years of experience in teaching and passed less than the 4 core courses related to ELT major. Hereinafter we call these raters as NEW.
B. Experienced raters who had over two years of experience in rating and receiving rater training, and over 5 years of experience in teaching and passed all the four core courses plus at least 2 selective courses related to ELT major. Hereinafter we call these raters as OLD.

2.2.Instruments

2.2.1. Oral tasks

The elicitation of test takers' oral proficiency was done through the use of five different tasks including Description, Narration, Summarizing, Role-play and Exposition tasks. Task 1 (*Description Task*) is an independent-skill task which reflects test takers' personal experience or background knowledge to respond in a way that no input is provided for it. On the other hand, tasks 3 (*Summarizing Task*) and 4 (*Role-play Task*) reflect test takers' use of their listening skills to respond orally. In other words, the content for the response was provided for the test takers through listening _ short or long. For tasks 2 (*Narration Task*) and 5 (*Exposition Task*) the test takers are required to respond to pictorial prompts including sequences of pictures, graphs, figures and tables.

2.2.2. Scoring Rubric

Each test taker's task performance was assessed using the ETS (2001) analytic rating scale. In ETS (2001) scoring rubric, individual tasks are assessed using appropriate criteria including *fluency*,

grammar, *vocabulary*, *intelligibility*, *cohesion* and *comprehension*. Each of these criteria is accompanied by a set of 7 descriptors.

2.3. Procedure

2.3.1. Pre-training phase

Prior to collecting any data from the test takers, the raters' background questionnaire was given to the raters to fill out. The aim was to enable the researcher to classify them into the two groups of rating expertise i.e., inexperienced and experienced raters. In order to run the speaking tasks, the 200 test takers were divided randomly into two groups in a way that half of the students took part in each phase of the study, i.e., pre, post-training.

2.3.2. Rater training

After the pre-training scoring phase, the raters participated in a training session in which the speaking tasks and the rating scale were introduced and time was given to practice the instructed materials with some sample responses. In addition to the training sessions, feedback on previous ratings was provided to each rater individually in the second norming session. In this respect, the raters having Z-scores beyond ±2 were considered to have a significant bias and were reminded individually to mind the issue accordingly. With respect to feedback on raters' consistency, the raters having infit mean squares beyond the acceptable range of 0.6 to 1.4, as suggested by Wright and Linacre (1994), were considered as misfitting in a way that the raters with an infit mean square value below 0.6 as too consistent (overfit the model) and those with an infit mean square value of above 1.4 as inconsistent (underfit the model). Therefore, the raters were pointed out individually on the issue if they were identified as misfitting.

2.3.3. Post-training phase

Immediately after the training program, the oral tasks were once again run. As it was mentioned before in the pre-training data collection procedure, the second half of the test takers (including 100 students) was used from whom to elicit data.

2.4. Data Analysis

In order to investigate the research questions a pre-post method research design was adopted to investigate the raters' development in rating L2 speaking performance (Cohen, Manion & Morrison, 2007). Quantitative data were collected and analyzed with a Multifaceted Rasch Model (MFRM) during two scoring sessions for the six test facets including test takers, rater, rater group, task, rating criterion and test version and their interactions to investigate variations in rater behavior and rater biasedness. The scoring patterns of the two groups of raters (inexperienced & experienced) were investigated each time they scored test takers' oral performances. The quantitative data were compared (1) across the two rater groups to investigate the raters' ability cross-sectionally at each rating point, and (2) within each rater group to investigate the development of the raters' ability. 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.

3. Results

RQ1: Are the scale categories used in the study at the same level of difficulty?

Table 1 displays the average scores given by the raters of each group of expertise to test takers' performance in each of the six scale categories before the training program. The table shows that NEW raters were more lenient than OLD raters and consequently assigned higher scores than OLD raters.

Table 1
Descriptive Statistics of Scores Given by Raters to Test Takers' Performance on Each 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

Furthermore, in order to determine whether there is a significant difference in raters scoring of test takers' oral performance ability, a one-way ANOVA on the scale categories was conducted respectively (Wright & Linacre, 1994). Table 2 represents the one-way ANOVA results of the raters' scoring of test takers' oral performances on each category.

Table 2
One-way ANOVA of Raters' Scoring of Test Takers' Oral Performance Ability on Each 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 outcome of the table reflects that there is a significant mean difference with respect to raters' scoring of test takers' oral performance ability on each scale category at the pre-training phase. Besides, in order to further investigate where exactly the significant mean difference is located, a post hoc Scheffé test was run for a pairwise comparison of category means. The outcome showed that there is significant mean difference between all pairs of categories with respect to their scorings of test takers' oral performance ability at the pre-training phase except for the following pairs: Cohesion-Intelligibility (p=0.222), and Vocabulary-Grammar (p=0.890).

In order to assess the validity of the analytic descriptors, MFRM was used. MFRM is useful in rating scale validation to analyze sources of variation in scoring. Besides, bias analysis examines the systematic sub-pattern interaction between raters and the rating scales (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 is capable of calculating raters' biases for each category of the rating scale by comparing the expected and observed values in a set of data and then reporting the outcome in a form of residuals. Later on, through converting the residuals into Z-scores, the bias value is obtained. This Z-score shows any significant deviation from what is expected from that particular rater allowing for routine and acceptable score variation. A Z-score between ±2 is regarded as a rater's normal scoring behavior thus an acceptable range of biasedness (McNamara, 1996). Table 3 represents the difficulty measurement report for the rating scale scoring categories at the pre-training phase of the study.

The *first column (Scale category)* displays the rating scale used in this study. The *second column (scale difficulty)* represents the difficulty of scale categories from Cohesion, as the most severely scored category, (Difficulty logit: 0.79) to Grammar, as the least severely scored category, (Difficulty logit: -0.46).

The *third column (SE)* shows that 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 referred to as *"quality control fit statistics"* which demonstrates to what extent the data fit the Rasch model, or in other words the difference between the observed scores and the expected ones. An observed score is the one given by a rater to a test taker, here, on a scale category, and an expected score is the one predicted by the model considering the facets involved (Wright & Linacre, 1994). In other words, Fit statistics typically is used to measure *within-rater consistency (Intra-rater consistency)* which shows to what extent each rater ranks the test takers consistent with his/her true ability. Fit statistics are classified into two subcategories entitled *infit* and *outfit* statistics. Infnit is the weighted mean square statistic which is weighted towards expected responses and thus *sensitive to unexpected responses* close to the point in which the decision is made. In other words, it is the mean score difference between actual scores and the estimated scores provided by the analysis. Outfit is the same as above but it is unweighted and is more *sensitive to sample size, outliers and extreme ratings* (Bonk & Ockey, 2003).

Fit statistics has an expected value of 1 and a range of zero to infinity; however, there is no definite range for interpreting fit statistics value or for setting the upper and lower limits; thus, the acceptability of fit is done on a judgmental basis not just on a statistical one. According to Myford and Wolfe (2004) such decisions are highly dependent to the assessment context. Wright and Linacre (1994) suggest an acceptable range within 0.6 to 1.4 logit values. Therefore, in order to examine the raters' fit statistics value, the researcher of this study employed it. 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 which are placed below this range are *overfit* or *too consistent and lack of variability*, showing that they were rated too consistent which indicates that the raters had difficulty separating the different scale categories, in other words, they do not use the whole scale category range, and those above this range are *underfit* **(misfit)** or *too inconsistent*, showing that they were rated too inconsistently. In literature, the terms underfit and overfit are often both referred to as misfit (Linacre,

2002). Here, Cohesion (Infit MnSq = 1.5) was identified as misfitting. This reflects that this category was rated too inconsistently by the raters before training.

However, the logit difficulty estimates do not alone tell us whether the differences in severity are meaningful or not; therefore, FACETS also provides us with several indications of the reliability of differences among the elements of each facet. The most helpful ones to study are ***Separation index***, ***Reliability*** and ***Fixed Chi-square*** which can be found at the bottom of the table.

The ***separation index*** is the measure of the spread of the estimates related to their precision. Adequate separation is important in situations in which a test produces scores that test users use to separate test takers into categories defined by their performance (Myford & Wolfe, 2004). In case the scale categories were equally difficult, the standard deviation of the scale category difficulty estimates should be equal to or smaller 1.00. Here, Cohesion was identified as the most severely scored category, (Difficulty logit: 0.79), and Grammar, as the least severely scored category, (Difficulty logit: -0.46), thus making the separation index of 1.25.

The ***reliability*** in the case of rating scale categories demonstrates the degree of agreement among raters in scale category difficulty. It shows to what extent or how well the analysis distinguishes among the various categories of the rating scale with respect to their difficulty in use by the raters. High values of rater separation reliability indicate significant differences among the rating scale categories. The high amount of *reliability index* in this phase ($r = 0.90$) indicates that the analysis could reliably separate the rating scale categories into various levels of difficulty.

Fixed Chi-square tests the null hypothesis to check whether all elements of the facet are equal or not. The Fixed Chi-square value for all the 6 rating scale categories was measured. The Chi-square value indicates whether there was a significant difference in rating scale categories level of difficulty ($X^2_{(5, N=6)} = 7464.12$, $p < 0.00$). Here, the high value of Chi-square indicates that at least two categories of the rating scale did not share the same on a parameter (e.g., difficulty). Consequently, the outcome suggested that the scale categories did not have the same level of difficulty.

Consequently, the *separation in the category difficulty* is rather high (1.25 logits) which shows that the rating scale categories were more than 1 statistically different levels of difficulty with a high *reliability of the separation index* (0.90). The high reliability shows that the scale categories were reliably separated with respect to their level of difficulty and that the analysis was reliable. As explained by Wink,

Gass, and Myford (2012) separation reliability indices close to zero show that scale categories did not differ significantly in terms of their levels of difficulty and that they had rather similar levels of difficulty; whereas the separation reliability indices close to 1.0 demonstrate that the scale categories were very reliably separated with respect to their difficulty levels. The *fixed Chi-square value* for all the 6 scale categories was measured ($X^2_{(5, N=6)} = 7464.12$, $p < 0.00$); therefore the null hypothesis that the scale categories were at the same level of difficulty 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 at 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 scores to test takers on Vocabulary and Grammar. The fact that the rating scale categories were rated having different difficulty measures indicate that the raters were not treating them the same way when rating. This means that the raters were able to discriminate among them.

Table 3
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 (all same) Chi-square: 7464.12, *df*= 5, *p*<0.00			
Scale category separation index: 1.25			
Reliability index: 0.90			

A bias analysis was run to investigate the rater-scale category interaction. Z-scores reveal to what extent raters had bias in their ratings. Bias is the difference between expected and observed ratings of the obtained data which is then divided by its standard error to achieve then Z-score (Linacre, 2002). The outcome of Z-score analysis shows the extent and direction of the bias. The Z-scores are also plotted into

a graph (Figure 2) showing raters' biasedness on each category of the rating scale map for each rater at each phase of the study.

Table 4 displays the *significant* rater-scale category interaction for the 20 raters. Since there were 6 categories in the analytic rating scale, 120 interactions were achieved. Among these 69 significant biases were observed of which 34 were positive (showing severity) and 35 were negative (showing leniency).

The *first column (Rater ID)* refers to the raters who had significant bias towards the scale categories presented in the *second column (Scale category)*.

The *third column (Obs-Exp average score)* displays the total observed score for all the 100 test takers participating at the pre-training phase on each scale category minus the total expected score for the students on each category. Since the possible score for each scale category falls in between 1 to 7; therefore, the total observed score for the test takers fall in between 100 to 700.

The *fourth column (bias logit)* displays the bias logit for each rater on the scale categories and the *fifth column (SE)* displays the error of the bias estimate which is quite low here, between 0.02 and 0.05 logits, showing the precision of measurement of raters' biases.

The *sixth column (Z-score)* displays bias estimates of biases converted into Z-scores. Z-scores between ±2.0 are considered as acceptable; thus a Z-score below -2.0 indicates that the rater consistently scored the scale category more leniently compared to the way that particular rater rates other categories. In contrast, a Z-score greater than +2.0 shows that the rater consistently scores the category more severely than other categories. There were 16 significant interactions in Cohesion; 16 in Intelligibility; 15 in Fluency; 8 in Comprehension; 8 in Vocabulary and 6 in Grammar. The number of significant interactions in the categories of the rating scale for each rater ranged from 1 to 6. The mean number of bias interactions for each rater was 3.45. With respect to the least and most biases towards the scale categories, one rater (OLD3) had only one bias interaction and 4 raters (NEW2, NEW6, OLD8 and OLD4) had six bias interactions. Raters had the tendency to show more severity biasedness towards Cohesion and Intelligibility each with 9 cases and more leniency towards Fluency with 8 cases.

Columns seven and *eight (Fit statistics)* display the infit mean square value which reflect how consistent the pattern of bias is for the rater to evaluate a particular scale category across all the test takers. This indicates which rater-scale category interaction was identified as misfitting. Regarding fit

statistics, a number of rater-category interactions were identified as misfitting, i.e., they were not placed within the acceptable range of 0.6 and 1.4 as suggested by Wright and Linacre (1994). The following raters were spotted as overfitting the model:

Table 4
Rater-Scale Category Significant Bias Interaction Analysis Report (Pre-training)

Raters ID	Scale category	Obs-Exp average score	Bias (logits)	SE	Z-score	Fit statistics	
						Infit Mn Sq	Outfit Mn Sq
OLD8	1	-0.35	0.68	0.05	3.24	1.0	1.0
OLD8	2	-0.44	0.84	0.04	4.06	0.4	0.4
OLD8	3	-0.26	0.51	0.04	2.43	0.8	0.8
OLD8	4	0.28	-0.53	0.04	-2.53	0.8	0.8
OLD8	5	0.36	-0.70	0.04	-3.35	0.6	0.7
OLD8	6	-0.39	0.74	0.04	3.55	0.5	0.5
OLD4	1	-0.32	0.61	0.04	2.94	0.9	0.9
OLD4	2	-0.28	0.53	0.02	2.53	0.8	0.8
OLD4	3	0.31	-0.59	0.04	-2.84	0.9	0.9
OLD4	4	-0.34	0.65	0.03	3.14	1.0	1.1
OLD4	5	0.50	-0.95	0.04	-4.56	0.5	0.5
OLD4	6	-0.39	0.74	0.04	3.55	1.1	1.1
OLD1	1	-0.32	0.61	0.04	2.94	0.9	0.9
OLD1	2	-0.53	1.01	0.03	4.87	1.3	1.3
OLD1	3	-0.50	0.95	0.04	4.56	1.2	1.2
OLD1	5	0.30	-0.57	0.04	-2.74	0.9	1.0
OLD1	6	-0.34	0.65	0.03	3.14	1.0	1.0
NEW3	1	-0.63	1.20	0.04	5.78	1.8	1.8
NEW3	2	-0.29	0.55	0.04	2.64	0.8	0.8
NEW3	3	-0.25	0.51	0.04	2.59	0.7	0.8
NEW3	5	-0.70	1.35	0.03	6.49	1.9	1.9
NEW7	1	-0.59	1.14	0.03	5.47	1.8	1.8
NEW7	3	-0.35	0.68	0.04	3.24	1.0	1.0
NEW7	5	0.34	-0.65	0.04	-3.14	1.0	1.0
OLD5	2	-0.44	0.84	0.04	4.06	1.3	1.3
OLD5	3	-0.53	1.01	0.04	4.87	1.6	1.6
OLD5	6	0.36	-0.70	0.04	-3.35	1.1	1.1
OLD10	3	-0.52	0.99	0.04	4.76	1.5	1.5
OLD10	4	-0.44	0.84	0.04	4.06	1.3	1.3
NEW4	1	0.36	-0.70	0.04	-3.35	1.1	1.1
NEW4	2	-0.33	0.63	0.04	3.04	1.0	1.0
OLD3	1	-0.92	1.77	0.05	8.52	0.7	0.8
OLD6	2	-0.34	0.65	0.04	3.14	1.0	1.0
OLD6	4	0.45	-0.87	0.05	-4.16	1.3	1.3
NEW9	1	0.47	-0.91	0.04	-4.36	1.4	1.4
NEW9	2	-0.55	1.06	0.04	5.07	1.6	1.5

NEW9	3	0.31	-0.59	0.04	-2.84	0.9	0.9
NEW1	1	0.39	-0.74	0.04	-3.55	1.1	1.1
NEW1	2	-0.43	0.82	0.03	3.95	1.3	1.4
NEW1	4	0.41	-0.78	0.02	-3.75	1.2	1.2
OLD2	1	-0.37	0.68	0.02	3.09	1.0	1.0
OLD2	2	0.33	-0.63	0.04	-3.04	1.0	1.1
OLD2	3	0.64	-1.22	0.04	-5.88	1.9	1.9
OLD9	1	0.51	-0.97	0.04	-4.66	1.5	1.5
OLD9	3	-0.35	0.68	0.04	3.24	1.0	1.0
NEW8	2	0.32	-0.61	0.03	-2.94	0.9	0.9
NEW8	3	0.51	-0.97	0.03	-4.66	1.5	1.4
NEW5	1	-0.31	0.59	0.04	2.84	0.9	0.9
NEW5	2	0.37	-0.72	0.04	-3.45	1.1	1.1
NEW5	3	0.61	-1.16	0.04	-5.58	1.8	1.8
NEW5	5	0.41	-0.78	0.04	-3.75	1.2	1.3
OLD7	1	-0.45	0.87	0.04	4.16	1.3	1.3
OLD7	2	0.26	-0.51	0.05	-2.43	0.8	.8
OLD7	4	0.29	-0.56	0.04	-2.47	0.8	.8
NEW10	1	0.31	-0.59	0.03	-2.84	0.9	.9
NEW10	2	0.34	-0.65	0.04	-3.14	1.0	1.0
NEW10	3	0.52	-1.00	0.03	-4.78	1.5	1.5
NEW2	1	0.70	-1.35	0.04	-6.49	1.8	1.8
NEW2	2	0.42	-0.80	0.05	-3.85	1.2	1.3
NEW2	3	0.50	-0.95	0.03	-4.56	1.5	1.5
NEW2	4	-0.52	0.99	0.04	4.76	1.5	1.5
NEW2	5	0.72	-1.37	0.04	-6.59	1.6	1.6
NEW2	6	-0.64	1.22	0.05	5.88	1.9	1.9
NEW6	1	0.32	-0.61	0.04	-2.94	0.9	0.9
NEW6	2	0.41	-0.78	0.03	-3.75	1.2	1.2
NEW6	3	0.54	-1.03	0.04	-4.97	1.6	1.7
NEW6	4	0.83	-1.58	0.04	-7.60	1.7	1.7
NEW6	5	-0.67	1.29	0.04	6.18	1.5	1.5
NEW6	6	0.42	-0.80	0.05	-3.85	1.2	1.2
Mean		0.00	0.00	0.03	0.00	1.16	1.18
SD		0.46	0.88	0.00	4.25	0.36	0.36

Chi-square: 845.64, *df*=119, *p*<0.00

1:Cohesion / 2:Intelligebility / 3:Fluency / 4:Comprehension / 5:Vocabulary / 6:Grammar

Rater OLD8 in both Comprehension and Vocabulary (infit Mn Sq. = 0.4 and 0.5); OLD4 in Vocabulary (infit Mn Sq. = 0.5). This shows that these raters overfitted the model and they were too consistent and relatively assigned similar scores on these scale categories. On the other hand, NEW3 in Cohesion and Vocabulary (infit Mn Sq. = 1.8 and 1.9); NEW7 in Cohesion (infit Mn Sq. = 1.8); OLD10 in Fluency (infit Mn Sq. = 1.5); NEW9 in Intelligibility (infit Mn Sq. = 1.6); OLD2 in Fluency (infit Mn

Sq. = 1.9); OLD9 in Cohesion (infit Mn Sq. = 1.5); NEW8 in Fluency (infit Mn Sq. = 1.5); NEW5 in Fluency (infit Mn Sq. = 1.8); NEW10 in Fluency (infit Mn Sq. = 1.5); NEW2 in Cohesion, Fluency, Comprehension, Vocabulary and Grammar (infit Mn Sq. = 1.8, 1.5, 1.5, 1.6 and 1.9); NEW6 in Fluency, Comprehension and Vocabulary (infit Mn Sq. = 1.6, 1.7 and 1.5). These raters underfitted (misfitted) the model and assigned rather inconsistent scores.

In order to better demonstrate the systematic pattern of rater-scale category bias interaction, Table 5 represents the frequency of bias interactions for each rating scale category in an ordered pattern. Below each category name, the frequency of severity or leniency biasedness is provided.

Table 5
Frequency of Rater-Category Bias Interaction (Pre-training)

Category➔ Rater🡻	Cohesion	Intelligibility	Fluency	Comprehension	Vocabulary	Grammar	Total
NEW1	L	S		L			1S / 2L
NEW2	L	L	L	S	L	S	2S / 4L
NEW3	S	S	S		S		4S / 0L
NEW4	L	S					1S / 1L
NEW5	S	L	L		L		1S / 3L
NEW6	L	L	L	L	S	L	1S / 5L
NEW7	S		S		L		2S / 1L
NEW8		L	L				0S / 2L
NEW9	L	S	L				1S / 2L
NEW10	L	L	L				0S / 3L
OLD1	S	S	S		L	S	4S / 1L
OLD2	S	L	L				1S / 2L
OLD3	S						1S / 0L
OLD4	S	S	L	S	L	S	4S / 2L
OLD5		S	S			L	2S / 1L
OLD6		S		L			1S / 1L
OLD7	S	L		L			1S / 2L
OLD8	S	S	S	L	L	S	4S / 2L
OLD9	L		S				1S / 1L
OLD10			S	S			2S / 0L
Total	9S / 7L	9S / 7L	7S / 8L	3S / 5L	2S / 6L	4S / 2L	34S / 35L

S: Severity
L: Leniency
$X^2_{(1)} = 4.52$, p=0.034

9 raters had severity and 7 leniency to *Cohesion*; 9 raters had severity and 7 leniency to *Intelligibility*; 7 raters had severity and 8 leniency to *Fluency*; 3 raters had severity and 5 leniency to *Comprehension*; 2 raters had severity and 6 leniency to *Vocabulary*; and finally, 4 raters had severity and 2 leniency to *Grammar*. In this table (Table 5), for example, rater NEW1 showed severity in Intelligibility

and leniency in both Cohesion and Comprehension, while rater OLD1 showed leniency in Vocabulary and severity in Cohesion, Intelligibility, Fluency and Grammar.

A Chi-square test was run to investigate any significant difference between the raters' interaction with the 6 scale categories. The result displayed a significant difference ($X^2_{(1)} = 4.52$, p=0.034) regarding the interaction between raters' scoring with any of the rating scale categories. This finding shows that raters had a substantial interactional difference in terms of showing significant bias to any of the scale rating categories. The higher frequency of rater-scale category bias interaction for Cohesion and Intelligibility (each with 16 cases) compared to Comprehension, Vocabulary and Grammar (8, 8 and 6 cases respectively) might indicate that the descriptors for the first two categories (Cohesion and Intelligibility) were somehow more difficult to agree compared to the others.

Table 6 represents the average scores given by the raters of each group of expertise to test takers' performance of each of the six rating scale categories used at the post-training phase. The table, like before, shows that NEW raters were more lenient than OLD raters and consequently assigned higher scores.

Table 6
Descriptive Statistics of Scores Given by Raters to Test Takers' Performance on Each Scale Category (Post-training)

Tasks	N	Mean			SD
		NEW	OLD	Both	(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, in order to determine whether there is a significant difference in raters scoring of test takers' oral performance ability after training, a one-way ANOVA on the scale categories was conducted. Table 7 represents the one-way ANOVA results of the raters' scoring of test takers' oral performance on each category at the post-training phase.

Table 7
One-way ANOVA of Raters' Scoring of Test Takers' Oral Performance Ability 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 outcome of the table reflects that there is a significant mean difference with regard to raters' scoring of test takers' oral performance ability on each scale category at the post-training phase. The outcome of the post hoc Scheffé test showed that there is 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 8 represents the difficulty measurement report for the rating scale scoring categories at 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 is rather high (0.83 logits) and the *reliability of the separation index* was high (0.87). The *fixed Chi-square* value for all the 6 scale categories was measured ($X^2_{(5,\ N=6)} = 811.63$, $p < 0.00$); therefore, the null hypothesis that the scale categories were at the same level of difficulty was rejected. In other words, the finding shows that there is a significant variation among the six rating scale categories regarding difficulty at the post-training phase. Through making a comparison between the two phases of the study, the range of scale category 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 tended to be harsher (less tolerant of weaknesses) on Cohesion, Intelligibility, Fluency and Comprehension, whereas they tended to assign higher scores to test takers on Grammar and Vocabulary.

The *fourth column (Infit MnSq.)* shows that, after training, no category was identified misfitting (beyond the acceptable 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 at the post-training phase.

Table 8
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 (all same) chi-square: 811.63, *df*= 5, *p*<0.00			
Scale category separation index: 0.83			
Reliability index: 0.87			

RQ2: Is there a reduction of rater biases in relation to the scoring of each rating scale category following training?

A bias analysis was run to investigate the rater-scale category interaction after training. Table 9 represents the *significant* rater-scale category interaction for the 20 raters. Since there were 6 categories in the analytic rating scale, 120 interactions were achieved. Among these 51 significant biases were observed of which 27 were positive (showing severity) and 24 were negative (showing leniency). There were 11 significant interactions in Cohesion; 12 in Intelligibility; 7 in Fluency; 4 in Comprehension; 11 in Vocabulary and 6 in Grammar and the number of significant interactions in the categories of the rating scale for each rater ranged from 0 to 6. The mean number of bias interactions for each rater was 2.55. Regarding the least and most biases towards the scale categories, one rater (NEW8) had no bias interaction and 1 rater (OLD8) had six bias interactions. Raters had tendency to show more severity biasedness towards *Cohesion* and *Intelligibility* each with 8 cases and more leniency towards Vocabulary with 8 cases.

For *Columns seven* and *eight (Fit statistics)* which display the infit mean square, the outcomes at the post-training phase demonstrated that rater OLD8 in Intelligibility was identified as overfitting the model (infit Mn Sq. = 1.5) and no rater was spotted as underfitting or misfitting the model.

Through making a comparison between the obtained data in the two phases of the study, the outcome demonstrated that the training program was effective enough in bringing the raters into the acceptable range of consistency. In other words, raters displayed a considerable extent of consistency

within one another in rating the test takers' oral performances regarding the use of rating scale categories. Once again this proves the positive effectiveness of the training program that raters were identified neither inconsistent nor too consistent.

Table 9
Rater-scale Category Significant Bias Interaction Analysis Report (Post-training)

Raters ID	Scale category	Obs-Exp average score	Bias (logits)	SE	Z-score	Fit statistics	
						Infit Mn Sq	Infit Mn Sq
OLD8	1	-0.31	0.58	0.04	2.77	0.9	0.9
OLD8	2	-0.38	0.73	0.03	3.46	1.5	1.6
OLD8	3	-0.23	0.43	0.02	2.04	0.7	0.8
OLD8	4	-0.25	0.47	0.04	2.23	0.7	0.7
OLD8	5	0.31	-0.59	0.04	-2.81	0.9	1.0
OLD8	6	-0.33	0.62	0.03	2.98	1.0	1.0
OLD4	1	-0.27	0.52	0.02	2.47	0.8	0.8
OLD4	2	-0.24	0.45	0.05	2.13	0.7	0.8
OLD4	3	-0.26	0.50	0.04	2.39	0.8	0.8
OLD4	4	-0.29	0.55	0.05	2.64	0.8	0.8
OLD4	5	-0.42	0.80	0.05	3.83	1.2	1.3
OLD7	1	-0.33	0.62	0.04	2.98	1.0	1.0
OLD7	2	-0.29	0.54	0.03	2.59	0.8	0.9
OLD7	3	-0.45	0.86	0.03	4.09	1.3	1.3
OLD7	5	0.42	-0.80	0.02	-3.83	1.2	1.3
OLD6	1	-0.26	0.50	0.04	2.39	0.8	0.8
OLD6	2	-0.29	0.55	0.05	2.64	0.8	0.9
OLD6	3	-0.43	0.81	0.03	3.86	1.2	1.2
NEW7	1	-0.33	0.64	0.05	3.03	1.0	1.0
NEW7	2	-0.27	0.51	0.04	2.41	0.8	0.9
OLD1	1	-0.49	0.93	0.06	4.45	1.4	1.4
OLD1	5	0.40	-0.75	0.04	-3.59	1.1	1.1
NEW3	2	-0.30	0.57	0.04	2.74	0.9	1.0
OLD10	5	-0.30	0.56	0.03	2.69	0.9	0.9
NEW5	3	0.37	-0.70	0.02	-3.35	1.1	1.1
NEW9	6	0.44	-0.84	0.02	-4.02	1.3	1.3
OLD5	6	-0.31	0.58	0.06	2.78	0.9	0.9
OLD5	2	0.42	-0.80	0.04	-3.81	1.2	1.4
NEW1	2	-0.38	0.71	0.04	3.41	1.1	1.2
NEW1	6	0.31	-0.60	0.04	-2.85	0.9	0.9
OLD2	1	-0.31	0.58	0.03	2.77	0.9	0.9
OLD2	4	0.46	-0.87	0.04	-4.16	1.3	1.3
NEW10	4	0.29	-0.55	0.04	-2.64	0.8	0.9
NEW10	5	0.36	-0.69	0.04	-3.29	1.1	1.3

NEW2	5	-0.40	0.77	0.03	3.66	1.2	1.3
NEW2	6	0.36	-0.68	0.03	-3.26	1.0	1.0
OLD9	1	-0.26	0.49	0.04	2.33	0.7	0.7
OLD9	2	0.31	-.59	0.04	-2.80	0.9	1.0
OLD9	5	0.37	-0.70	0.05	-3.32	1.1	1.1
NEW4	1	0.35	-0.66	0.04	-3.15	1.0	1.0
NEW4	2	0.24	-0.46	0.04	-2.18	0.7	0.8
NEW4	5	0.25	-0.48	0.03	-2.27	0.7	0.7
NEW6	1	0.32	-0.62	0.06	-2.94	0.9	1.0
NEW6	2	-0.40	0.76	0.06	3.61	1.2	1.2
NEW6	3	0.28	-0.54	0.04	-2.58	0.8	0.8
NEW6	5	0.27	-0.52	0.05	-2.47	0.8	0.9
OLD3	1	0.43	-0.82	0.03	-3.91	1.3	1.3
OLD3	2	0.26	-0.50	0.05	-2.40	0.8	1.0
OLD3	3	0.30	-0.57	0.06	-2.74	0.9	0.9
OLD3	5	0.39	-0.73	0.06	-3.49	1.1	1.2
OLD3	6	0.35	-0.66	0.04	-3.15	1.0	1.1
Mean		-0.01	0.01	0.03	0.08	0.97	1.02
SD		0.34	0.65	0.01	3.11	0.20	0.21

Chi-square: 363.74, *df*=119, *p*<0.00

1:Cohesion / 2:Intelligebility / 3:Fluency / 4:Comprehension / 5:Vocabulary / :Grammar

It is noteworthy to note that the existence of various levels of difficulty and bias measures on the rating scale categories could make one hypothesize the existence of halo effect which indicates that the rating scale categories are conceptually treated as different by raters (Myford & Wolfe, 2004). In order to test whether the observed difficulty and bias variations in rating scale categories is the reflection of halo effect or the effect of rating categories, a non-significant all fixed Chi-square will indicate the existence of halo effect (Myford & Wolfe, 2004). However, the existence of a *significant* Chi-square, both before and after training, signifies that the raters did not have any halo effect with respect to the scoring of rating scale categories. Besides, the high reliability index showed that the raters had variation with regard to their conceptual treatment. Consequently, little halo effect could be expected.

In order to better demonstrate the systematic pattern of rater-scale category bias interaction, Table 10 represents the frequency of bias interactions for each rating scale category. Below each category name, the frequency of severity or leniency biasedness is provided. After training, 8 raters had severity and 3 leniency in Cohesion; 8 raters had severity and 4 leniency in Intelligibility; 4 raters had severity and 3 leniency in Fluency; 2 raters had severity and 2 leniency in Comprehension; 3 raters had severity and 8 leniency in Vocabulary; and finally, 2 raters had severity and 4 leniency in Grammar. In the following

table, for example, rater NEW1 showed severity in Intelligibility and leniency in Grammar, while rater OLD1 showed severity in Cohesion, and leniency in Vocabulary.

Table 10
Frequency of rater-category bias interaction (Post-training)

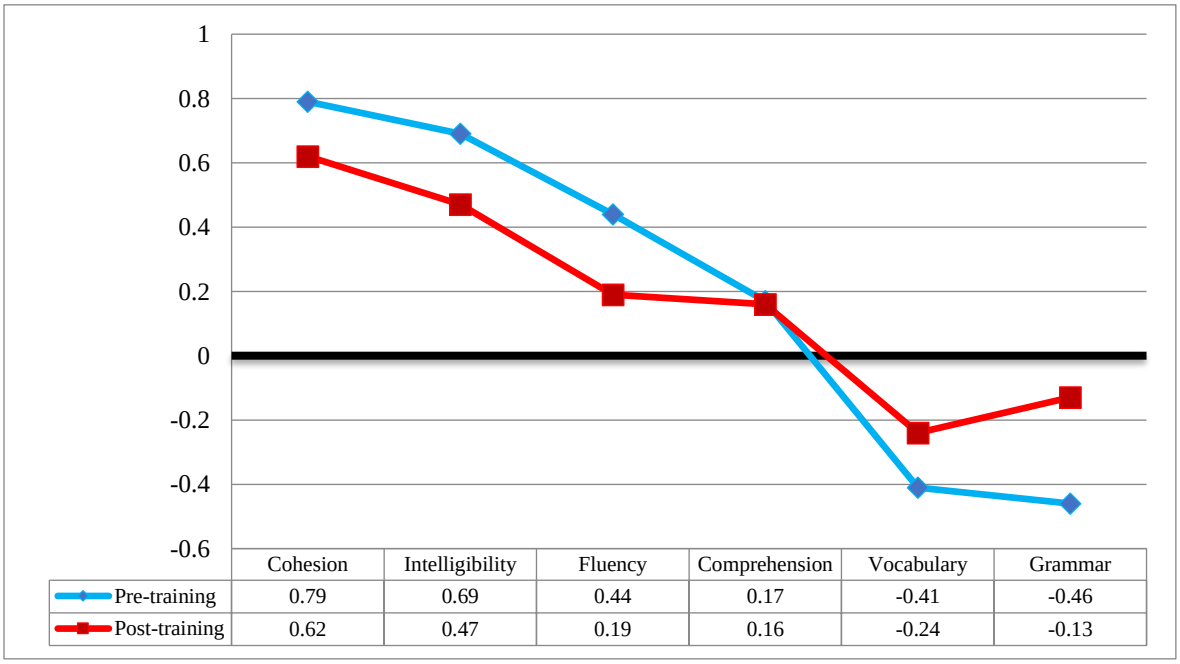
Category➔ Rater🡻	Cohesion	Intelligibility	Fluency	Comprehension	Vocabulary	Grammar	Total
NEW1		S				L	1S / 1L
NEW2					S	L	1S / 1L
NEW3		S					1S / 0L
NEW4	L	L			L		0S / 3L
NEW5			L				0S / 1L
NEW6	L	S	L		L		1S / 3L
NEW7	S	S					2S / 0L
NEW8							0S / 0L
NEW9						L	0S / 1L
NEW10				L	L		0S / 2L
OLD1	S				L		1S / 1L
OLD2	S			L			1S / 1L
OLD3	L	L	L		L	L	0S / 5L
OLD4	S	S	S	S	S		5S / 0L
OLD5		L				S	1S / 1L
OLD6	S	S	S				3S / 0L
OLD7	S	S	S		L		3S / 1L
OLD8	S	S	S	S	L	S	5S / 1L
OLD9	S	L			L		1S / 2L
OLD10					S		1S / 0L
Total	8S / 3L	8S / 4L	4S / 3L	2S / 2L	3S / 8L	2S / 4L	27S / 24L

S: Severity
L: Leniency
$X^2_{(1)} = 2.273$, p=0.132

Likewise the previous phase, a Chi-square test was run to investigate any significant difference between the raters' interaction with the 6 scale categories. At the post-training phase, the outcome showed no significant difference ($X^2_{(1)} = 2.273$, p=0.132) regarding the interaction between raters' scoring with any of the rating scale categories. This finding shows that the raters did not have any substantial interactional difference in terms of showing significant bias to any of the scale rating categories. The number of rater-scale category bias interactions is a lot fewer than the pre-training phase considerably. Intelligibility, Cohesion and Vocabulary were identified as the most controversial

categories for the raters to agree on, whereas Comprehension, Fluency and Grammar were found to be the least difficult ones for them to agree on. As already mentioned, the need for further investigation in future research with respect to the above-mentioned outcome which might be attributed to the ambiguity of descriptors used for each of the categories is felt.

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.



	Cohesion	Intelligibility	Fluency	Comprehension	Vocabulary	Grammar
Pre-training	0.79	0.69	0.44	0.17	-0.41	-0.46
Post-training	0.62	0.47	0.19	0.16	-0.24	-0.13

Figure 1. Scale category difficulty measure before and after training

RQ3: How does each rater's interactional bias to each rating scale category change throughout the study?

Additionally, in order to have a more precise picture of raters' biases in each category of the rating scale, Figure 2 plots the graphical representation of average rater-scale category significant bias interactions. The average of significant bias interaction between raters and scale categories was measured through adding up the significant biases for each category of the scale divided by the number of raters. The scale categories are placed on the horizontal axis and the Z-scores on the vertical axis. It shows to what extent raters had biases towards the rating scale categories.

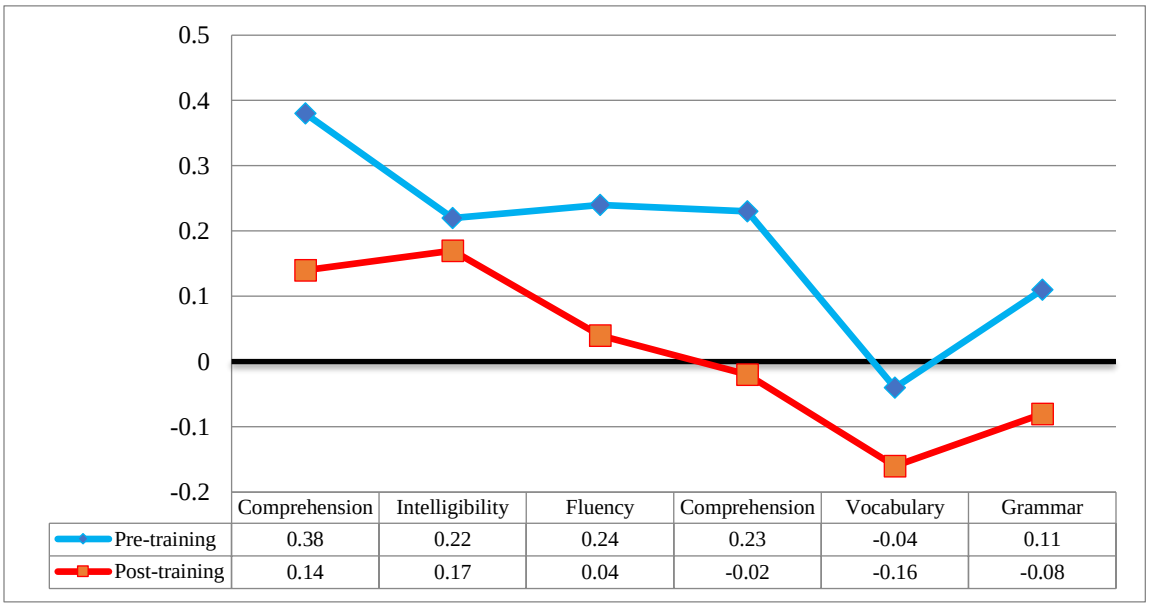


Figure 2. Rater-scale category average significant bias interaction before and after training

Table 11
Rater-Scale Category Biasedness (Pre-training)

Rater	Cohesion		Intelligibility		Fluency		Comprehension		Vocabulary		Grammar	
	Logit	Z	Logit	Z	Logit	Z	Logit	Z	Logit	Z	Logit	Z
NEW1	-0.74	**-3.53**	0.82	**3.91**	-0.28	-1.33	-0.44	**-2.09**	-0.37	-1.76	-0.22	-1.05
NEW2	-1.35	**-6.44**	-0.80	**-3.82**	-0.95	**-4.53**	0.99	**4.72**	-1.37	**-6.53**	1.22	**5.82**
NEW3	1.20	**5.72**	0.55	**2.62**	0.51	**2.43**	0.39	1.86	1.35	6.44	0.18	0.86
NEW4	-0.70	**-3.34**	0.63	**3.01**	0.08	0.38	0.14	0.67	-0.17	-0.81	0.21	1.00
NEW5	0.59	**2.81**	-0.72	**-3.43**	-1.16	**-5.53**	-0.35	-1.67	-0.78	**-3.72**	-0.31	-1.47
NEW6	-0.61	**-2.91**	-0.78	**-3.72**	-1.03	**-4.91**	-1.58	**-7.54**	1.29	**6.15**	-0.80	**-3.82**
NEW7	**0.47**	**2.24**	0.38	1.81	0.68	**3.24**	0.24	1.14	-0.65	**-3.10**	-0.11	-0.52
NEW8	0.13	0.62	-0.61	**-2.91**	-0.97	**-4.63**	-0.28	-1.34	-0.41	-1.96	-0.37	-1.76
NEW9	-0.91	**-4.34**	1.06	**5.06**	-0.59	**-2.81**	-0.24	-1.14	0.14	0.67	-0.12	-0.57
NEW10	-0.59	**-2.81**	-0.65	**-3.10**	-1.00	**-4.77**	0.16	0.76	-0.39	-1.86	-0.40	-1.90
OLD1	0.61	**2.91**	1.01	**4.82**	0.95	**4.53**	0.20	0.95	-0.57	**-2.72**	0.65	**3.10**
OLD2	0.68	**3.24**	-0.63	**-3.01**	-1.22	**-5.82**	-0.37	-1.76	-0.04	-0.19	0.10	0.48
OLD3	1.77	**8.44**	-0.11	-0.52	0.08	0.38	-0.04	-0.19	-0.19	-0.91	-0.22	-1.05
OLD4	0.61	**2.91**	0.53	**2.53**	-0.59	**-2.81**	0.65	**3.10**	-0.95	**-4.53**	0.74	**3.53**
OLD5	0.28	1.34	0.84	**4.01**	1.01	**4.82**	0.32	1.53	0.37	1.76	-0.7	**-3.34**
OLD6	-0.16	-0.76	0.65	**3.10**	-0.24	-1.14	-0.87	**-4.15**	-0.13	-0.62	0.19	0.91
OLD7	0.87	**4.15**	-0.51	**-2.43**	-0.40	-1.90	-0.56	**-2.67**	-0.41	-1.96	-0.39	-1.86
OLD8	0.68	**3.24**	0.84	**4.01**	0.51	**2.43**	-0.53	**-2.53**	-0.70	**-3.34**	0.74	**3.53**
OLD9	-0.97	**-4.63**	-0.40	-1.91	0.68	**3.24**	-0.38	-1.81	0.07	0.33	-0.32	-1.53
OLD10	0.27	1.29	0.23	1.10	0.99	**4.72**	0.84	**4.01**	0.26	1.24	0.31	1.48

Bold numbers represent significant biases

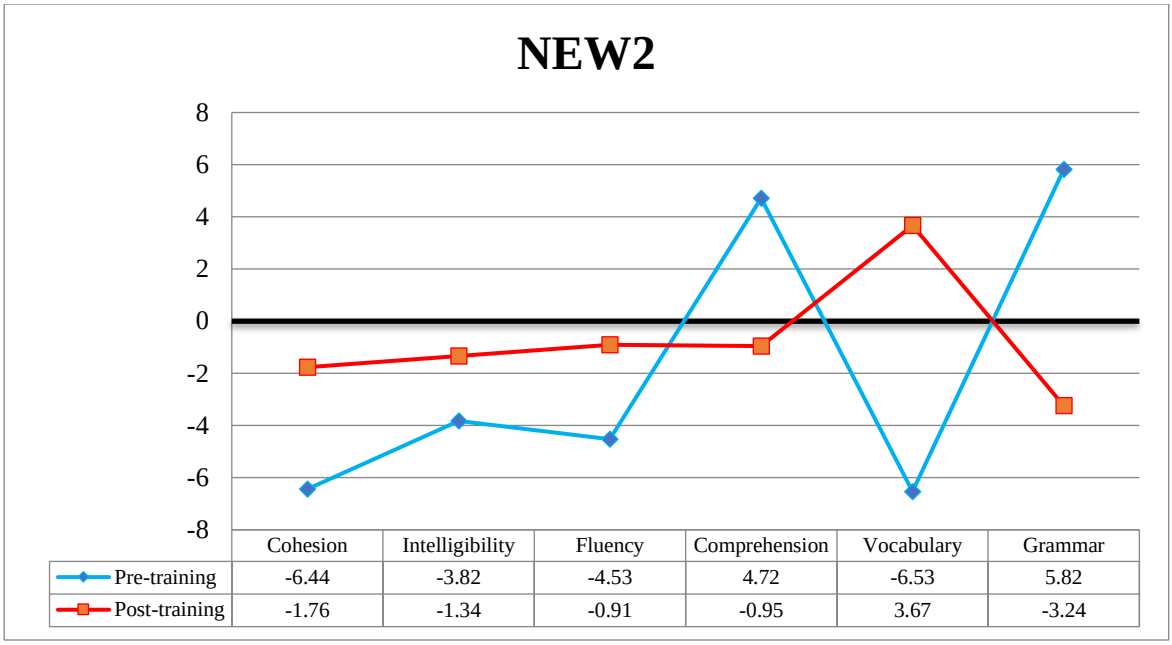
Tables 11 and 12 display each rater's biasedness to the categories of the rating scale before and after training. It should be indicated that unlike the above diagram, the following tables display both significant and non-significant rater-scale category biases.

Table 12
Rater-Scale Category Biasedness (Post-training)

Rater	Cohesion		Intelligibility		Fluency		Comprehension		Vocabulary		Grammar	
	Logit	Z	Logit	Z	Logit	Z	Logit	Z	Logit	Z	Logit	Z
NEW1	-0.28	-1.34	0.71	**3.39**	-0.32	-1.53	-0.19	-0.91	-0.25	-1.19	-0.60	**-2.86**
NEW2	-0.37	-1.76	-0.28	-1.34	-0.19	-0.91	-0.20	-0.95	0.77	**3.67**	-0.68	**-3.24**
NEW3	0.17	0.81	0.57	**2.72**	0.19	0.91	0.16	0.76	-0.11	-0.52	0.08	0.38
NEW4	-0.66	**-3.15**	-0.46	**-2.19**	-0.31	-1.48	-0.29	-1.38	-0.48	**-2.29**	0.12	0.57
NEW5	0.32	1.53	0.09	0.43	-0.70	**-3.34**	0.17	0.81	0.19	0.91	-0.14	-0.67
NEW6	-0.62	**-2.96**	0.76	**3.63**	-0.54	**-2.58**	-0.22	-1.05	-0.52	**-2.48**	-0.34	-1.62
NEW7	0.64	**3.05**	0.51	**2.43**	0.21	1.00	0.26	1.24	0.15	0.72	0.37	1.76
NEW8	0.11	0.52	0.06	0.29	0.17	0.81	-0.07	-0.33	-0.04	-0.19	0.08	0.38
NEW9	0.16	0.76	0.18	0.86	0.13	0.62	-0.10	-0.48	-0.18	-0.86	-0.84	**-4.01**
NEW10	-0.08	-0.38	0.17	0.81	-0.25	-1.19	-0.55	**-2.62**	-0.69	**-3.29**	-0.27	-1.29
OLD1	0.93	**4.44**	0.24	1.14	0.19	0.91	0.16	0.76	-0.75	**-3.58**	0.09	0.43
OLD2	0.58	**2.77**	-0.34	-1.62	-0.27	-1.29	-0.87	**-4.15**	-0.26	-1.24	-0.18	-0.86
OLD3	-0.82	**-3.91**	-0.50	**-2.39**	-0.57	**-2.72**	-0.38	-1.81	-0.73	**-3.48**	-0.66	**-3.15**
OLD4	0.52	**2.48**	0.45	**2.15**	0.50	**2.39**	0.55	**2.62**	0.80	**3.82**	0.21	1.00
OLD5	-0.17	-0.81	-0.80	**-3.82**	-0.16	-0.76	-0.23	-1.10	0.11	0.52	0.58	**2.77**
OLD6	0.50	**2.39**	0.55	**2.62**	0.81	**3.86**	0.27	1.29	0.33	1.57	-0.14	-0.67
OLD7	0.62	**2.96**	0.54	**2.58**	0.86	**4.10**	0.32	1.53	-0.80	**-3.82**	0.25	1.19
OLD8	0.58	**2.77**	0.73	**3.48**	0.43	**2.05**	0.47	**2.24**	-0.59	**-2.81**	0.62	**2.96**
OLD9	0.49	**2.34**	-0.59	**-2.81**	-0.25	-1.19	-0.26	-1.24	-0.70	**-3.34**	0.13	0.62
OLD10	-0.17	-0.81	-0.13	-0.62	0.07	0.33	0.12	0.57	0.56	**2.67**	-0.08	-0.38

Bold numbers represent significant biases

Figure 3 displays the assessment map which graphically plots the analysis of rater-scale category bias interaction (e.g., Rater NEW2) showing the bias interaction of the rater to the rating scale categories both before and after training.



	Cohesion	Intelligibility	Fluency	Comprehension	Vocabulary	Grammar
Pre-training	-6.44	-3.82	-4.53	4.72	-6.53	5.82
Post-training	-1.76	-1.34	-0.91	-0.95	3.67	-3.24

Figure 3. Rater NEW2 bias in the two phases of the study

Rater NEW2 showed significant bias in all of the scale categories at the pre-training phase. Her ratings on Comprehension and Grammar were severe, whereas for the remaining categories she was substantially lenient. At the post-training phase, the rater still showed extreme biased in two categories, i.e., Vocabulary and Grammar. Surprisingly, although this rater showed significant leniency in Vocabulary and severity in Grammar at the pre-training phase, he tended to switch her behavior thus showing respectively extreme severity and leniency after training. The reason might be due to the fact that the rater had an extreme interpretation of the feedback given for rating this category.

RQ4: To what extent are the rating scale categories similar or dissimilar to each other in rating and how does training affect their interrelationship?

In order to determine where each scale category at each phase of the study is located, a Multidimensional Scaling (MS) analysis on scale categories was performed. The analysis demonstrates the relative similarity/dissimilarity of each category in the form of its distance in relation to the other categories. The MS was done on the basis of considering the extent of rater-scale category interactional indices. The

following figures and tables (Figure 4 and 5 and Tables 13 to 14) display the MS analysis of the scale categories in the two phases of the study.

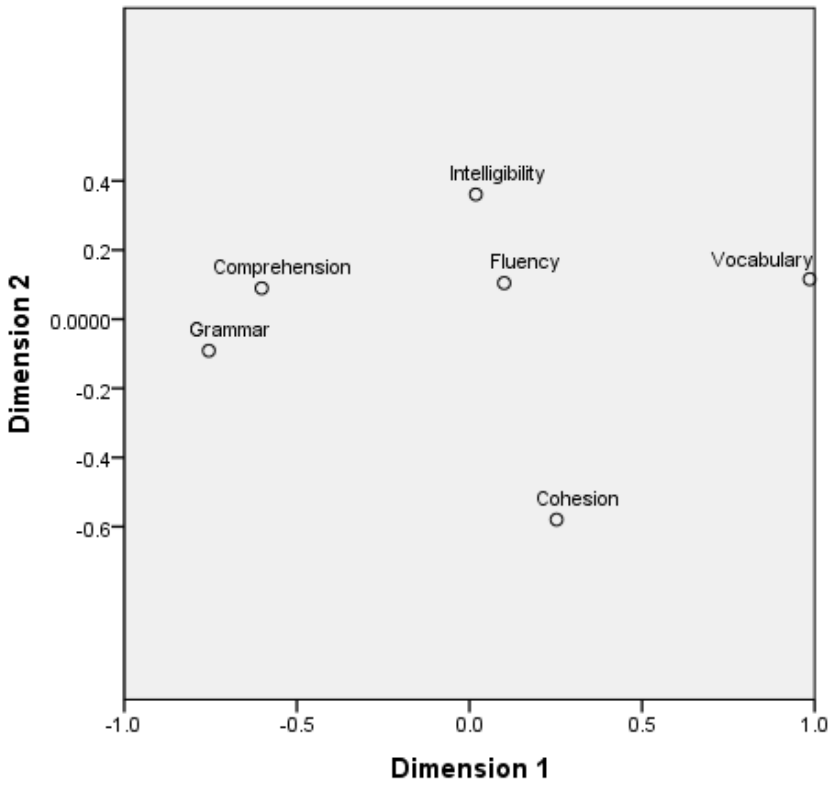


Figure 4. Multidimensional scaling for scale categories (Pre-training)

Table 13
Multidimensional Scaling Statistics of Scale Categories (Pre-training)

	Cohesion	Intelligibility	Fluency	Comprehension	Vocabulary	Grammar
Cohesion	0.000					
Intelligibility	0.293	0.000				
Fluency	0.319	0.026	0.000			
Comprehension	0.565	0.858	0.884	0.000		
Vocabulary	1.138	0.845	0.819	1.703	0.000	
Grammar	0.722	1.015	1.041	.157	1.860	0.000

At the pre-training phase, the most similarity was found between Fluency and Intelligibility tasks (Distance = 0.026), and the least one between Grammar and Vocabulary (Distance = 1.860).

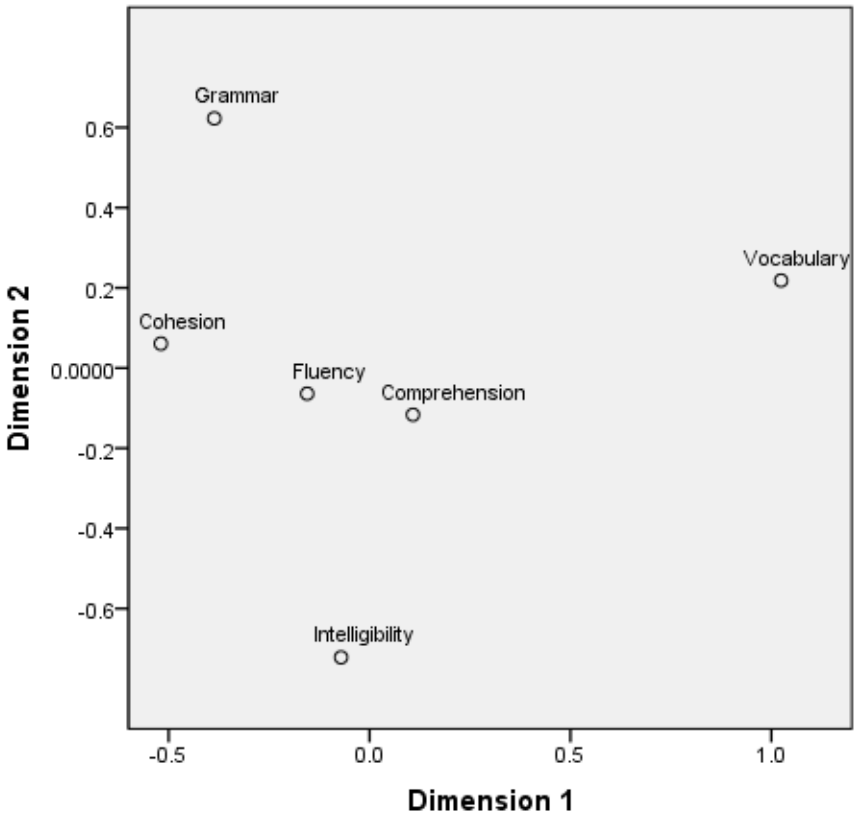


Figure 5. Multidimensional scaling for scale categories (Post-training)

Table 14
Multidimensional Scaling Statistics of Scale Categories (Post-training)

	Cohesion	Intelligibility	Fluency	Comprehension	Vocabulary	Grammar
Cohesion	0.000					
Intelligibility	0.164	0.000				
Fluency	0.497	0.333	0.000			
Comprehension	0.908	0.744	0.411	0.000		
Vocabulary	1.805	1.641	1.308	0.897	0.000	
Grammar	0.788	0.624	0.291	0.120	1.017	0.000

At the post-training phase, there is less distance variation among the categories in dimension one compared to the pre-training phase; however, in dimension two, raters tended to move a little farther away (around 0.2 distance range). At this phase, the most similarity was found between Grammar and Comprehension (Distance = 0.120), and the least one between Vocabulary and Cohesion (Distance = 1.805).

A pairwise comparison of category distance reflected that the following pairs reduced distance: Intelligibility-Cohesion (from 0.293 pre-training to 0.164 post-training); Comprehension-Intelligibility (from 0.858 pre-training to 0.744 post-training); Comprehension-Fluency (from 0.884 pre-training to 0.411 post-training); Vocabulary-Comprehension (from 1.703 pre-training to 0.897 post-training); Grammar-Intelligibility (from 1.015 pre-training to 0.624 post-training); Grammar-Fluency (from 1.041 pre-training to 0.291 post-training); Grammar-Comprehension (from 0.157 pre-training to 0.120 post-training); and Grammar-Vocabulary (from 1.860 pre-training to 1.017 post-training). However, some other pairs increased distance: Fluency-Cohesion (from 0.319 pre-training phase to 0.497 post-training); Fluency-Intelligibility (from 0.026 pre-training phase to 0.333 post-training); Vocabulary-Cohesion (from 1.138 pre-training phase to 1.805 post-training); Vocabulary-Intelligibility (from 0.845 pre-training to 1.641 post-training); Vocabulary-Fluency (from 0.819 pre-training to 1.308 post-training); Comprehension-Cohesion (from 0.565 pre-training to 0.908 post-training); and Grammar-Cohesion (from 0.722 pre-training to 0.788 post-training).

4. Discussion and Conclusion

The fact that the raters had a great deal of severity/leniency in rating using the rating scale categories, specifically at the pre-training phase, might most probably refer to their variability in interpreting the meaning of each criteria and its relevant descriptor. Still relatively similar outcome of rater training was observed which showed raters' substantial differences which even training could not be effective enough to eliminate them. The outcome of the study indicated that the raters improved consistency and reduced severity/leniency and bias after the training program on account of the use of the rating scale categories. The remaining differences regarding bias measures could probably be attributed to the result of 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 finding is fairly consistent with that of (Bijani & Fahim, 2011; Lumley & McNamara, 1995; Fulcher, Davidson & Kamp, 2011; 1993; Winke, Gass & Myford, 2012).

The outcomes of this study indicated that MFRM can be used to investigate raters' scoring behavior and can result in enhancement in rater training and validating the functionality of the rating scale descriptors. This results in the use of rating scale consistently by raters. MFRM data analysis let us identify which categories of the rating scale could well be discriminated by the raters. That is, MFRM

can be used to assist test developers whether the marking of scale descriptors are employed the way they were intended (McNamara & Knoch, 2012). MFRM can point out to sources of raters' bias thus making assessment fairer. It can lessen the intimidation of being either accepted or rejected based on the factors which have nothing to do with their true ability. Besides, it can determine raters' bias showing the extent to which raters show interaction to the categories of the rating scale. This can help provide feedback to help raters using rating scales in a more consistent was. The result of fit statistics at the post-training phase indicated that the raters were able to use the rating scale descriptors in a consistent way to score test takers' oral performances despite the various observed levels of severity.

With respect to the analytic rating scale used in the study, 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 have demonstrated that raters show halo effect when they face problems using rating scales. For example, Sawaki (2007) stated that when raters cannot identify certain aspects of rating scales, they resort to more global and holistic use of rating scales. This phenomenon was also observable in this study as well; however, training program proved to be effective enough in reducing this halo effect providing raters with more explicit scoring criteria. In other words, training helped raters use the descriptors of the rating scale 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 finding showed that halo effect is not necessarily only as a result of rater effect but can also be as a result of rating scale. However, no matter what kind it is, the effect can be reduced, although not neutralized totally by rater training programs.

In general, the differences in difficulty levels of the scale categories and fit statistics indicate that raters were able to discriminate the various categories of the rating scale. The results displayed that the rating scale descriptors are not stable in a way that they change radically with respect to raters' scoring from before training to after training. This outcome supports the variable competence model of second language acquisition by Tarone (1983) which indicates that individual's understanding of language is variable. It must also be indicated that this point of view is also confirmed in language testing by Skehan (1987, cited in Bachman, 2004).

The outcome of the study showed that training can result in higher levels of interrater consistency and reduced levels of severity/leniency, biasedness and inconsistency. However, it cannot turn rater into

duplicates of one another, but to make them more self-consistent. In other words, training cannot easily eradicate raters' individual differences related to their characteristics. Also, some amount of severity was still left after training which may have an impact on future interpretations and decisions. 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 be due to the rating scale used in the study. Perhaps the use of a valid rating scale has benefitted raters (both experienced and inexperienced) achieve consistency in scoring. Thus, it is suggested that similar research be conducted with a holistic rating scale or even without a rating scale to observe the possible contribution of a rating scale in training.

References

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.

Bazyar M. (2023). Process types in Persian-speaking aphasic discourse: A systemic functional approach. *Iranian Journal of Applied Linguistics, 26*(1). http://ijal.khu.ac.ir/article-1-3203-en.html

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.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110.

Bridley, G. (1998). Describing language development: Rating scales and SLA. In L. F. Bachman & A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112-140). Cambridge: Cambridge University Press.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33.

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

Fulcher, G., Davidson, F., & Kamp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29.

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

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

Knoch, U. (2007). *Diagnostic assessment of writing: The development and validation of a rating scale*. Unpublished PhD dissertation, University of Auckland.

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

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

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

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

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

Maldar M. (2022). The effect of online assessment on speaking complexity, accuracy, and fluency of Iranian intermediate EFL learners. Iranian Journal of Applied Linguistics, 25(1), 61-77. http://ijal.khu.ac.ir/article-1-3191-en.html

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

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.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555-576.

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.

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33-56.

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.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

Tarone, E. (1983). On the variability of interlanguage systems. *Applied Linguistics, 4*(2), 142-164.

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.

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.

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