Journal of Modern Research in English Language Studies
Vol. 5, No. 4, 2018, 27-53

Effectiveness of a Face-to-Face Training Program on Oral Performance Assessment: The Analysis of Tasks Using the Multifaceted Rasch Analysis

Houman Bijani[1*]

[1*] Assistant Professor, Department of English Language Teaching, Zanjan Branch, Islamic Azad University, Zanjan, Iran, *houman.bijani@gmail.com*

Abstract

The current popularity of second/foreign language oral performance assessment has led to a growing interest in tasks as a tool for assessing language learners' oral abilities. However, most oral assessment studies so far have investigated tasks separately; therefore, any possible relationship among them has remained unexplored. Twenty English as a foreign language (EFL) teachers rated the oral performances produced by 200 EFL learners before and after a rater training program using description, narration, summarizing, role-play, and exposition tasks. The findings demonstrated the usefulness of multifaceted Rasch measurement (MFRM) in detecting rater effects and demonstrating the consistency and variability in rater behavior aiming to evaluate the quality of rating. The outcomes indicated that test difficulty identification is complex, difficult, and at the same time multidimensional. On the other hand, test takers' ability is a more determining factor in their score variation than other intervening variables. The outcomes displayed no relationship between task difficulty and raters' interrater reliability measures. The findings suggest that tasks have various effects on oral performance assessment tests and most importantly, performance conditions in estimating the oral ability of test takers. Since various groups of raters have biases to different tasks in use, the findings indicated that training programs can reduce raters' biases and increase their consistency measures. The findings imply that decision makers had better not be concerned about raters' expertise in oral assessment, whereas they should establish better rater training programs for raters to increase assessment reliability.

Keywords: Bias; Interrater Consistency; Multifaceted Rasch Measurement (MFRM); Oral Proficiency Task; Severity/Leniency

Received 05 March 2019	Accepted 20 June 2019
Available online 08 July 2019	DOI: 10.30479/jmrels.2019.10667.1335

1. Introduction

Task, according to O'Sullivan (2002), is defined as "bounded classroom activities in which learners use language communicatively to achieve an outcome with the overall purpose of learning language" (p. 278). The current popularity of performance assessment has led to a growing interest in the tasks as a tool for assessing learner ability. Task-based assessment engages students in the performance of tasks, which stimulates the kind of language found in the real world situation with the purpose of eliciting authentic language sample from the test takers.

One issue regarding variation in test takers' performances attributes to task characteristics. This variation results in different scores under various conditions; thus, making it a feature of interest for further investigation. In the area of second language acquisition (SLA) research, the classification of tasks for the sake of better understanding of their influence on test takers' performances goes back to the early 1980s (Kyle, Crossley & McNamara, 2016). When discussing the scoring tasks, Skehan and Foster (1999) suggest that the task developer should consider the complexity and length of any texts which are to be used, the difficulty of the vocabulary needed to complete the task, the expected speed of speech, the number of speakers, the explicitness of information, discourse structure, and the amount of non-linguistic support available. On the part of the test taker, they further added that, the task developer should be aware of the level of confidence in using the language, motivation, prior learning experience, ability level, culture knowledge and awareness, and linguistic knowledge.

The appearance of item response theory (IRT) has made it possible to investigate task difficulty as in isolation from rater severity (Winke, Gass & Myford, 2012). This is based on the assumption that the scores awarded to an individual on a speaking task are influenced by his/her speaking proficiency, difficulty of the task, and the severity of the rater(s). In fact, very little is known about task difficulty or the difficulty of various tasks as they are compared with one another. Consequently, one of the most important challenges, which influences task characteristics, is how to determine task difficulty. This can help us in the appropriate use of task ranges which will clarify the way levels of performance are described. Some scholars (e.g., Robinson, 2001; Skehan, 1998) have identified a number of factors that affect task choice. As an example, Robinson (2001) identified two factors which have contribution to task difficulty: "resource directing factor" which deals with the number of task elements, their reasoning and immediacy of information provided, and "resource depleting factor" which deals with time planning, the number of tasks and prior knowledge. Robinson further claimed that by the manipulation of these factors, task performance will vary resulting in variation of task quality.

2. Literature Review

2.1. Factors Affecting Learners' Second Language Task Production

The chief concern of oral performance assessments is to evaluate test takers' substantial speaking ability obtained from their performance. However, in the course of assessment, test takers' oral performance is influenced by a number of factors other than their speaking ability. In this respect, Fulcher, Davidson and Kamp (2011) emphasized that test takers are not separated figures who are only responsible for their performance, whereas the interaction among factors such as interlocutors, raters, and test methods also influence test takers' performance. According to Ling, Mollaun and Xi (2014) one of the most substantial factors influencing test takers' oral performance is the nature of oral tasks used in assessment which include genre of the tasks, task type, task structure, task condition, and the level of cognitive complexity of the tasks which influence the oral performance of second language learners. As Trace, Janssen and Meier (2017) claim, the L2 learner`s oral performance differs from task to task. So, L2 learner`s oral productions will be different when they perform different oral task types, and consequently these different types of oral tasks will result in variation, called "task-induced variation". May (2009) agrees with this variation and asserts that in performing different tasks, learner`s oral production of some grammatical, morphological and phonological forms will vary in a particular manner. Skehan and Foster (1999) investigated the role of task type in foreign language oral production in terms of accuracy, fluency, and complexity. Two types of tasks (instruction task and an argumentative task) were used in the study and the outcome showed that the participants in the instruction-task group performed significantly better than those in the argumentative-task group in terms of accuracy, fluency, and complexity. The argumentative speeches were produced with more complex language than the instruction ones; however, fluency was higher in instruction speeches. In terms of accuracy, instruction-task group performed better than those in argumentative-task group, but argumentative speeches were more accurate than instruction speeches.

2.2. The Effect of Task Type on Oral Production

Oral assessment is often carried out by considering students' ability to produce words and phrases by evaluating their ability in doing a variety of tasks such as asking and answering questions about themselves, doing role-plays, making up minidialogues, defining or talking about some pictures or talking about given theme. As Robinson (2001) stated, features of second language oral output such as accuracy, fluency and complexity differ by task type. These three aspects, complexity, accuracy, and fluency of learners' performance are considered as learners' language ability determining factors

(Robinson, 2001). Studies incorporating task have been primarily concerned with analyzing the impact of task design on the accuracy, fluency and complexity of language in oral production.

In'nami and Koizumi (2016) suggested field testing of tasks along with the use of questionnaires to elicit test takers' and raters' perceptions to determine good and poor tasks. In an attempt, they scaled a number of oral speaking tasks, used in the ACTFL oral assessment, based on their functions. Then, by the use of a Rasch partial credit model, they assessed the difficulty of a number of tasks. Although they found a reasonable correlation between the suggested difficulty level and the assessment of difficulty by raters, this is far from testing tasks on students and assessing task difficulty from scores.

On behalf of language testing and assessment, Elder, Iwashita and McNamara (2002) claim that the more difficult and complex a task is, the more difficult it will be. In an attempt, they aimed to modify Skehan's (1998) model of task difficulty factors through investigating the following criteria: 1) *Perspective:* To tell a story from one's own perspective or from a third person's, 2) *Immediacy:* To tell a story with and without pictures, 3) *Adequacy:* To tell a story from a complete set of pictures, and with two or four missing ones, and 4) *Planning time:* To do an oral task with 2-3 minutes planning time, and without it. Nakatsuhara (2011) investigated the effect of planning time on test takers' accuracy, fluency, and complexity measures. The study revealed that those test takers who had planning time had a better performance with respect to complexity (number of subordinations), fluency (number of self-repairs), and accuracy (lack of grammatical mistakes). Ahmadian and Tavakoli (2011) investigated the effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in the oral production of EFL learners in the context of Iran. The results obtained from one-way ANOVAs revealed that the opportunity to engage simultaneously in careful online planning and task repetition enhances accuracy, complexity, and fluency significantly. In another study Kuiken and Vedder (2014) aimed at investigating the impact of planning conditions on the oral performance of the EFL learners while performing structured vs. unstructured tasks. Results demonstrated that planning time served no impact with regard to the accuracy and fluency of the learners' performances, but resulted in more complex performances when participants conducted the unstructured task. In the meantime, task structure did not affect the accuracy and complexity whilst promoting the fluency under the planned condition. Davis (2016) discussed the fact that 1 min of pre-task planning should be considered as an alternative to extend the face validity of the test. Moreover, although pre-task instructions displayed some role for diverting attention to form, planning did not serve any impact. Leaper and Riazi (2014) hypothesized that the Here-and-Now/There-and-Then narrative would be

more complex than the other versions of the task. The results showed that learners who performed the most complex versions of the task were significantly less fluent, with no such large differences regarding either structural or lexical complexity, and with significant improvements with regard to error-free units but not target-like use of articles.

In speaking, rater training is used to modify raters' expectations of tasks and test takers' characteristics, and to clarify various elements of the rating scale in order to reduce levels of rater variability (Khabbazbashi, 2017). Training is used to reduce extreme differences by minimizing random errors between raters in terms of severity and to increase the self-consistency of individual raters by reducing random errors (Davis, 2016). Closely related to training are the concepts of rater experience or rater expertise. Because scoring second language oral proficiency is done by raters, they are an essential part of proficiency assessment. Therefore, not only does rating reflect test takers' oral ability, but also raters' assessment schemes (Attali, 2016). A variety of researches on experienced and inexperienced raters' performances have indicated higher inter-rater consistency following training (Attali, 2016; Bijani, 2010). Commonly, in a majority of studies, extremely severe or lenient inexperienced raters have benefited from the training program thus have modified their rating behavior making it like the other raters'. In a study by Bijani (2010) on the effect of rater training on rater consistency scoring test takers' written language proficiency, the consistency of inexperienced raters improved much more after training compared to experienced raters.

However, the relative contribution of task factors to the success of any given task is mostly unknown. Although it is frequently claimed that lack of specialist knowledge about the task topic makes the task difficult for test takers, there is little evidence in this case. Measures of task difficulty, bias and consistency measures have not been investigated precisely so far. On the other hand, almost all studies so far have investigated tasks separately; therefore, any possible relationship among them has remained unexplored. Besides, the notion of task difficulty and its relationship to underlying subcategory measures of fluency, accuracy and complexity in various task conditions has not been addressed comprehensively and there is little evidence suggesting which tasks are more suitable for test takers' of particular ability levels. Also, few, if any, studies have used a pre- and post-training design in their investigations of task difficulty measures in relation to other intervening facets. Therefore, in order to answer the above-mentioned questions, the following research questions were formed:

1. Is there a reduction of rater biases with respect to the tasks of various difficulty levels following the training program?

2. Is there any significant difference in rating tasks with various difficulty measures before and after training?

3. Does test takers' score variability reflect their true speaking ability?

4. Is there any significant relationship between task difficulty and raters' interrater reliability in scoring?

3. Method

3.1. Participants

Two hundred adult Iranian students of English as a Foreign Language (EFL), including 100 males and 100 females participated as test takers. The students were selected from intermediate, high-intermediate, and advanced levels studying at the Iran Language Institute (ILI).

twenty Iranian EFL teachers, including 10 males and 10 females participated in this study as raters. These raters were undergraduate and graduate in English language related fields of study, teaching in different universities and language institutes. In order to fulfill the requirements of this study, the raters were 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 based on 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.

3.2. Instruments

3.2.1. The Speaking Test

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*) was an independent-skill task which reflected 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*) reflected test takers' use of their listening skills to respond orally. In tasks 2 (*Narration Task*) and 5 (*Exposition Task*) the test takers were required to respond to pictorial prompts including sequences of pictures, graphs and tables. The tasks were implemented via two methods of task delivery: (1) direct and (2) semi-direct. The direct test was designed for use in an individual face-to-face method (i.e., speaking to an interlocutor _ here a rater), whereas the semi-direct test was designed for use in a language laboratory setting.

For the purpose of comparability, both formats of the test consisted of one-way exchanges (monologic) in which the test takers were required to communicate information in response to prompts from the interviewer/rater. However, on the direct version of the test, the role play allowed for a more authentic information gap activity in which meaning was negotiated between a test taker and an interviewer (dialogic). In terms of their structure, the tasks used in this study were characterized on the basis of the model developed by Gardner (1992). In the first place, the tasks were classified as either planned (allowing preparation time) or unplanned (designed to elicit spontaneous language). In the third place, they were distinguished as either open (allowing a range of possible solutions) or closed (allowing a restricted set of possible responses). In the fourth place, the tasks were also classified as being convergent (involving problem-solving in which the aim was to arrive at a particular goal) and those which were divergent (without specific goals, involving decision making, opinion and agreement). In this study, the only two-way task, role-play, was regarded to be convergent. In another classification, tasks were classified with respect to perspective dimension. This was to ask the test takers to do the tasks from their own (first person perspective) or another person's point of view (third person perspective). Finally, tasks were classified with regard to their immediacy dimension. This was to ask the test takers to speak using Here-and-now and There-and-then language structures.

In this respect, the task types used in this study could be classified into two categories with respect to their difficulty level on the basis of the given factors above (Robinson, 2001). The following table (Table 1) gives the classification of tasks and their predicted difficulty on behalf of the given factors with respect to their difficulty levels.

3.2.2. The Scoring Rubric

For both versions of the test, 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*.

Table 1

Table of Predicted Task Difficulty Classification

Dimension	Difficult (predicted)	Easy (predicted)
Openness	Close (limited response)	Open (free response)
Information exchange direction	Dialogic	Monologic
Language convergence / divergence	Convergent	Divergent
Language planning	Without planning time	With planning time
Perspective	3rd person point of view	1st person point of view
Immediacy	There-and-then	Here-and-now

3.3. Procedures

3.3.1. Pre-training Phase

Raters' background questionnaire was given to the raters to fill out before starting to run the test tasks and collect data. The aim was to enable the researcher to classify them into the two groups of rating expertise i.e., inexperienced raters and experienced ones. The 200 test takers participating as data providers were divided randomly into two groups in a way that each group took part in each stage of the study (pre-, and post-training), and from each group half of the test takers took the direct and half the semi-direct test version.

3.3.2. Rater Training Program

After the pre-training phase, the raters participated in a training (norming) session in which the speaking tasks and the rating scale were introduced and time was given to practice the instructed material with some sample responses. The training program consisted of rater norming and feedback on previous rating behavior and was conducted in two separate norming sessions, each lasting for about six hours, with an interval of one week. It is noteworthy to indicate that each training session started with a brief warm-up of approximately 30 minutes in which the purpose of the study and the nature of the instruments were elaborated. It was reiterated that during the norming session the raters would learn about test tasks and scoring criteria. They would also have an opportunity to practice scoring several sample responses using the criteria. The sample performances used in the training program were selected among those representing a various range of test takers' oral proficiency levels. Moreover, the raters discussed differences in their scores and reviewed their decision making processes with the

instructor. A norming packet was used in the norming session including the tasks, representative samples of oral performances from previous ratings representing various scoring bands to better provide raters with awareness of the scoring principles, and the analytic scoring rubric; however, the packet did not include a transcription of sample responses to make it similar to the real rating sessions since during an actual rating session, raters score test takers' responses as they listen to them.

Regarding feedback on raters' biases, the raters who had z-scores beyond ±2 were considered to have a significant bias and were reminded individually to mind the issue. For feedback on raters' consistency, the raters who had 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.

3.3.3. Post-training Phase

After the training program, the tasks of both versions of the test were run. The second half of the test takers (including 100 students) was used from whom to elicit data. All the raters participating in this study were given one week to submit their scorings.

3.4. Data Analysis

In order to investigate the research questions, the researcher employed a pre-post, research design using a quantitative approach to investigate the raters' development with regard to rating L2 speaking performance (Cohen, Manion & Morrison, 2007). Quantitative data (i.e., raters' scores based on an analytic scoring rubric) were collected and analyzed with a Multifaceted Rasch Model (MFRM) during two scoring sessions including the facets of test takers, rater, rater group, task, rating criteria, test version, and their interactions to investigate variations in rater behavior and rater biasedness.

4. Results and Discussion

4.1. Results

1. Is there a reduction of rater biases with respect to the tasks of various difficulty levels following the training program?

2. Is there any significant difference in rating tasks with various difficulty measures before and after training?

Table 2 displays the average scores given by the raters of each group of expertise to test takers' performance in each of the five tasks before the training. The table shows that NEW raters were more lenient than OLD raters and consequently assigned higher scores than OLD raters.

Table 2

Descriptive Statistics of Scores Given by Raters to Test Takers' Performance on each Oral Task (Pre-training)

Tasks	N	Mean			SD (Both)
		NEW	OLD	Both	
Description	100	32.04	28.98	30.48	0.72
Narration	100	27.72	23.88	25.80	0.48
Summarizing	100	31.38	26.52	28.92	0.54
Role Play	100	28.80	24.42	26.58	0.16
Exposition	100	26.82	19.14	22.98	0.22
Mean		29.35	24.58	26.95	0.42
SD		2.27	3.64	2.89	0.23

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 task types was conducted (Steiger, 1980). Table 3 represents the one-way ANOVA results of the raters' scoring of test takers' oral performance on each task.

Table 3

One-way ANOVA of Raters' Scoring of Test Takers' Oral Performance Ability on each Task (Pre-training)

	Sum of Squares	*df*	Mean Square	F	Sig.
Between Groups	1995.572	4	498.893	33.166	0.000
Within Groups	7445.890	495	15.042		
Total	9441.462	499			

$p<0.05$

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 oral task 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 administered for a pairwise comparison of task means. The outcome displayed a significant mean difference among all pairs of tasks with respect to their scorings of test takers' oral performance ability at the pre-training phase except for narration-role play (p=0.671). To further investigate the raters' behavior, a FACETS analysis was also run to investigate the rater-task interactions. Data analysis, out of 10,000 (20×100×5) interactions at the pre-training phase, revealed the following

information. Table 4 demonstrates the task difficulty measures by rater groups and bias analysis between rater groups and tasks.

The *first column (Tasks)* displays the tasks used in the study. The *second column (Raw score average)* represents the raters' mean scores given to the test takers on each task. The highest scored task appears at the top (description, logit value: 2.42) and the lowest scored task appears at the bottom (exposition, logit value: 1.72). This is the average outcome of subtracting each task's expected given score from their observed given raw score.

The *third column (Fair average)* demonstrates the extent to which the mean ratings of task given scores differ. For instance, here, the mean rating of the most lenient scored task was 2.42 and the fair average was 2.83. Similarly, the mean rating of the severest rater was 1.72 and the fair average was 2.39. The data show that the two extreme scored tasks were 0.7 raw scores apart when comparing the mean ratings and 0.44 raw scores when comparing their fair averages. According to Winke, Gass and Myford (2012) both values demonstrate raters severity spread; however, the difference is that fair average is a better estimate when not all raters scored all the tasks.

Column four (Difficulty logit measure) shows that the exposition task was the most difficult that (difficulty logit = 0.82) and the description task was the least difficult task (difficulty logit = -0.37), thus making a spread of 1.19 logit range difference.

Columns five and *six (Bias logits)* demonstrate each rater group biasedness, their severity and leniency measures, to any of the tasks of the test. Further explanation will be provided later.

Column seven (SE) displays the standard error of estimation which was rather small, between 0.03 and 0.05, indicating the high precision of measurement. Therefore, the smaller the standard error, the more precise the ratings will be.

Column eight and *nine (infit and outfit mean square)* are referred to as *"quality control fit statistics"* which show 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 on one criterion for a task, and an expected score is the one predicted by the model considering the facets involved (Wright & Linacre, 1994). In other words, fit statistics simply is used to determine *within-rater consistency (Intra-rater consistency)* which indicates the extent to which each rater ranks the test takers consistent with his/her true ability. Infnit is the weighted mean square statistic which is weighted towards expected responses and thus *sensitive to unexpected responses* near the point where the decision is made.

Outfit is the same as above but it is unweighted and is more *sensitive to sample size, outliers and extreme ratings* (Eckes, 2015). Fit statistics has an expected value of one and a range of zero to infinity; however, there is no straightforward rule or universally definite range for interpreting fit statistics value; therefore, the acceptability of fit is done on a judgmental basis. The one which enjoys more popularity among the others is the one offered by Wright and Linacre (1994) which suggests an acceptable range within 0.6 to 1.4 logit values. Therefore, in order to investigate the raters' fit statistics value, the researcher of this study adopted it. The raters who are placed below this range are *overfit* or *too consistent and lack of variability* (showing that they do not use the whole scale category range and overuse certain categories of them), and those above this range are *underfit* (misfit) or *too inconsistent*, that is, they have too much variability and are different from the expected ratings than the model predicts. Here, summarizing (Infit MnSq = 0.4) was identified as overfitting indicating too consistency or lack of variability in scoring and exposition (Infit MnSq = 1.8) was identified as misfitting showing that it was rated inconsistently and with too much variability before training.

Column ten (r) displays the point biserial correlation which is the correlation coefficient between each task and the rest of the tasks rated in this study. In other words, it shows how similarly the tasks were scored by the raters. Values lower than 0.30 show the tasks whose ratings are not consistent with the ratings of the rest of the tasks. Here, the exposition task (correlation coefficient = 0.72) and narration task (correlation coefficient = 0.77) had the least correlation and the most correlation coefficient with the rest of the tasks respectively.

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 are *separation index*, *reliability* and f*ixed chi-square* which can be found below the table. The separation index is the measure of the spread of the estimates related to their precision. In other words, it is the ratio of the corrected standard deviation of element measures to the Root Mean Square Estimation Error (RMSE) which shows the number of statistically distinct levels of difficulty among the tasks. In case the tasks were equally difficult, the standard deviation of the task difficulty estimates should be equal to or smaller than the RMSE of the entire data set which results in a separation index of 1.00 or even less (if there is a total agreement among tasks in their difficulty, the separation index should be 0.00). In the case of this phase of the study, exposition was identified as the most severely scored category, (difficulty logit: 0.82), while description, as the least severely scored category, (difficulty logit: -0.37), thus making the

separation index of 1.19. The reliability in the case of rating tasks demonstrates the degree of agreement among raters in task difficulty scoring. It shows to what extent or how well the analysis distinguishes among the various tasks with respect to their difficulty in use by the raters. High values of rater separation reliability indicate significant differences among the rating tasks. The high amount of *reliability index* in this phase of the study (r = 0.91) indicates that the analysis could reliably separate the tasks 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 five tasks was measured. The chi-square value indicates whether there was a significant difference in tasks' level of difficulty ($X^2_{(4,\ N=5)}$ = 774.67, p<0.00). Here, the high value of chi-square indicates that at least two tasks did not share the same on a parameter (e.g., difficulty). Consequently, the outcome suggested that the tasks did not have the same level of difficulty.

Table 4

Task Difficulty and Bias Analysis Measure between Rater Groups and Tasks (Pre-training)

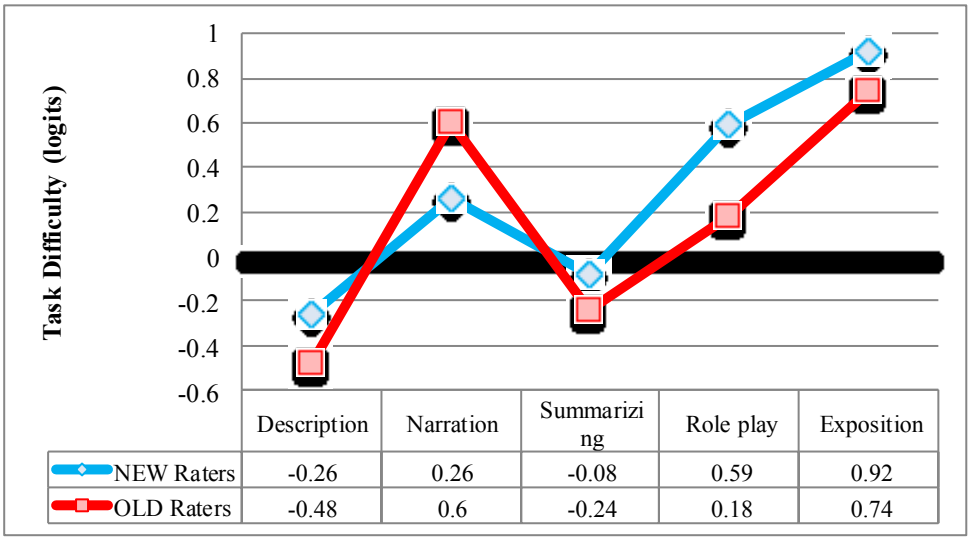
Tasks	Raw score average	Fair average	Difficulty rating measure (logits) (Both groups)	Bias logits (z –Score)		SE	Infit mean square	Outfit mean square	*r*
				NEW	OLD				
Description	2.42	2.83	-0.37	-1.20	-0.65	0.03	1.0	1.1	0.73
Narration	2.07	2.67	0.43	0.68	1.50	0.04	1.2	1.3	0.77
Summarizing	2.33	2.79	-0.16	-0.60	-0.20	0.03	0.4	0.5	0.75
Role play	2.16	2.74	0.39	1.47	0.45	0.04	1.1	1.1	0.75
Exposition	1.72	2.39	0.82	2.30	1.95	0.05	1.8	1.8	0.72
Mean	2.14	2.68	0.22	0.53	0.61	0.03	1.10	1.16	0.74
SD	0.27	0.17	0.48	1.44	1.10	0.00	0.50	0.46	0.02

Fixed (all same) chi-square: 774.67, df = 4, p<0.00
Task difficulty Separation index: 1.19
Reliability index: 0.91

Furthermore, in order to make sure whether there is a significant difference between NEW and OLD raters with regard to task rating difficulty for each particular task, an independent t-test was run. The obtained result, $t_{\text{Description}}$ (18) = 41.19, p<0.01; $t_{\text{Narration}}$ (18) = 52.27, p<0.01; $t_{\text{Summarizing}}$ (18) = 71.56, p<0.01; $t_{\text{Role Play}}$ (18) = 29.12, p<0.01; and $t_{\text{Exposition}}$ (18) = 56.68, p<0.01, showed that there is significant difference between NEW and OLD raters with respect to their difficulty level in rating.

Figure 1 demonstrates task difficulty derived from NEW and OLD rater groups. The logit task difficulty measure for NEW raters ranged from -0.26 (description) to 0.92 logits (exposition) making a whole logit spread of 1.18 logits. For OLD raters, the range of task difficulty measure was rather similar to NEW raters. The task logit range was from -0.48 logits (description) to 0.74 logits (exposition) making a whole logit spread of 1.22

logits. The figure also displays that both groups demonstrated relatively similar patterns in task difficulty measures. The description task was given the lowest difficulty measure and similarly the exposition task was given the highest difficulty measure by both rater groups.



	Description	Narration	Summarizing	Role play	Exposition
NEW Raters	-0.26	0.26	-0.08	0.59	0.92
OLD Raters	-0.48	0.6	-0.24	0.18	0.74

Figure 1. Task Difficulty Measures by NEW and OLD Rater Groups (Pre-training)

As it was already indicated, a bias analysis, *columns five* and *six*, was also run to investigate the interaction between rater groups and tasks. The extent to which rater groups were biased towards tasks was measured based on z-scores. In this respect, z-values between ±2 are regarded as the acceptable range of insignificant biasedness; thus, any value beyond this range is seen as significantly bias to tasks.

The bias analysis between raters and tasks rather confirmed the outcomes of the previous findings of the study in a way that some degree of significant bias was observed between tasks and rater groups. In this respect, NEW raters showed significant biasedness, in particular, too severity in exposition (logit value = 2.30). Moreover, OLD raters, although they were within the acceptable range of biasedness (logit value = 1.95), they were rather on the borderline of biasedness. At this phase of the study, the least degree of biasedness for NEW raters was in description (logit value = -1.20) and again for OLD raters in exposition (logit value = -0.65). This finding is rather confirmed by the outcomes of task difficulty measures in which Exposition task was identified the most difficult one. Figure 2 displays the bias analysis of the interaction between rater groups and tasks at the pre-training phase.

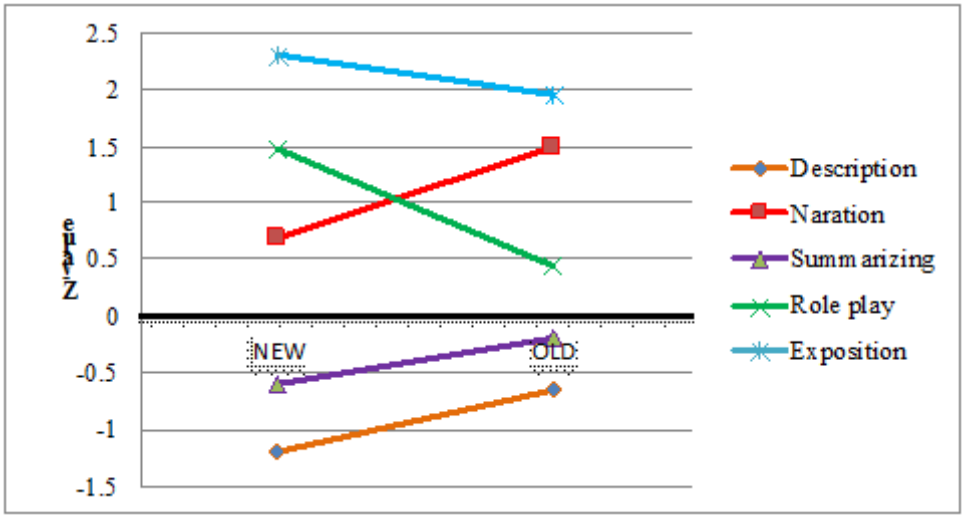


Figure 2. Rater-task Bias Interaction (Pre-training)

In summary, some bias was shown for all the tasks of the study by both groups of raters; however, significant severity bias was obtained by OLD raters for the exposition task.

The facility of most tasks (description; summarizing; role play, and exposition) for OLD raters to score compared to NEW raters could most probably be relevant to their overfamiliarity with such tasks in their previous ratings. The exception of narration task in the above list, and the reason that made it harder for OLD raters to score at the pre-training phase is not very clear. It is hypothesized that there might have been something with the test format or perhaps with the scoring rubric which made the task more difficult to score. However, the analysis of verbal protocols revealed that NEW raters expected high quality performances from the test takers and since they were not satisfied enough with their production, it was difficult for them to come up with the right scoring. Table 5 represents once again the average scores given by the raters of each group of expertise to test takers' performance in each of the five tasks used at the post-training phase. The table, similar to the pre-training phase, shows that NEW raters were more lenient than OLD raters and consequently assigned higher scores than OLD raters.

Furthermore, in order to determine whether there is significant difference in raters' scoring of test takers' oral performance ability after training, a one-way ANOVA on the task types was conducted (Steiger, 1980). Table 6 represents the one-way ANOVA results of the raters' scoring of test takers' oral performance on each task at the post-training phase.

Table 5

Descriptive Statistics of Scores Given by Raters to Test Takers' Performance on each Oral Task (Post-training)

Tasks	N	Mean			Sd. (Both)
		NEW	OLD	Both	
Description	100	32.58	30.66	31.62	0.44
Narration	100	30.12	27.60	28.86	0.30
Summarizing	100	31.86	30.42	31.14	0.32
Role Play	100	26.88	25.08	25.98	0.08
Exposition	100	24.84	23.64	24.24	0.12
Mean		29.25	27.48	28.36	0.25
SD		3.30	3.13	3.21	0.14

Table 6

One-way ANOVA of Raters' Scoring of Test Takers' Oral Performance Ability on each Task (Post-training)

	Sum of Squares	*df*	Mean Square	F	Sig.
Between Groups	3217.840	4	804.460	87.186	0.000
Within Groups	4567.360	495	9.227		
Total	7785.200	499			

$p<0.05$

The outcome of the table reflects that there is significant mean difference with respect to raters' scoring of test takers' oral performance ability on each oral task at the post-training phase. In addition, in order to further investigate where exactly the significant mean difference is located, a post hoc test with Scheffé procedure was run for a pairwise comparison of task means. The outcome of the table displays that there is significant mean difference among all pairs of tasks with respect to their scorings of test takers oral performance ability at the post-training phase except for the following pairs: description-summarizing (p = 0.194), and role play-summarizing (p = 0.142).

To further investigate the rater-oral task interaction effect, a FACETS analysis was also run to investigate the rater-task interactions. Data analysis, out of 10,000 (20×100×5) interactions at the post-training phase, revealed the following information. Table 7 demonstrates the task difficulty measures by rater groups and bias analysis between rater groups and tasks.

The *second column (Raw score average)* represents the raters' mean scores given to the test takers on each task. The highest scored task appears at the top (description, logit value: 2.78) and the lowest scored task appears at the bottom (exposition, logit value: 1.88). The *third column (Fair average)*, on the other hand, demonstrates the extent to which the mean ratings of task given scores differ.

Column four (difficulty logit measure) shows that the exposition task was the most difficult task (difficulty logit = 0.61) and the description task was the least difficult task (difficulty logit = -0.14), thus making a spread of 0.75 logit range difference.

Column seven (SE) displays the standard error of estimation which was rather small, between 0.04 and 0.06, indicating the high precision of measurement.

Column ten (r) displays the point biserial correlation after training. Here, the exposition task (correlation coefficient = 0.86) and summarizing task (correlation coefficient = 0.93) had the least correlation and the most correlation coefficient with the rest of the tasks respectively. Below the table, the task difficulty *separation index* measured 0.75 and the *reliability index* 0.84 showing that the analysis rather well distinguishes among various levels of task difficulty. The *fixed hi-square* value measured ($X^2_{(4,\ N=5)} = 229.52$, $p<0.00$), showing that there was a significant difference between the tasks with regard to their difficulty.

Similar to the pre-training phase, in order to make sure whether there is a significant difference between NEW and OLD raters with regard to rating difficulty of each particular task, an independent t-test was run. The result, $t_{\text{Description}}(18) = 1.86$, $p>0.05$; $t_{\text{Narration}}(18) = 0.87$, $p>0.05$; $t_{\text{Summarizing}}(18) = 1.33$, $p>0.05$; $t_{\text{Role Play}}(18) = 38.66$, $p<0.01$; $t_{\text{Exposition}}(18) = 21.47$, $p<0.01$, displayed no significant difference with respect to the task difficulty between NEW and OLD raters for description, narration and summarizing tasks was found. However, still there observed significant difference between the two groups of expertise for the remaining role play and exposition tasks with respect to their rating difficulty measures.

Table 7

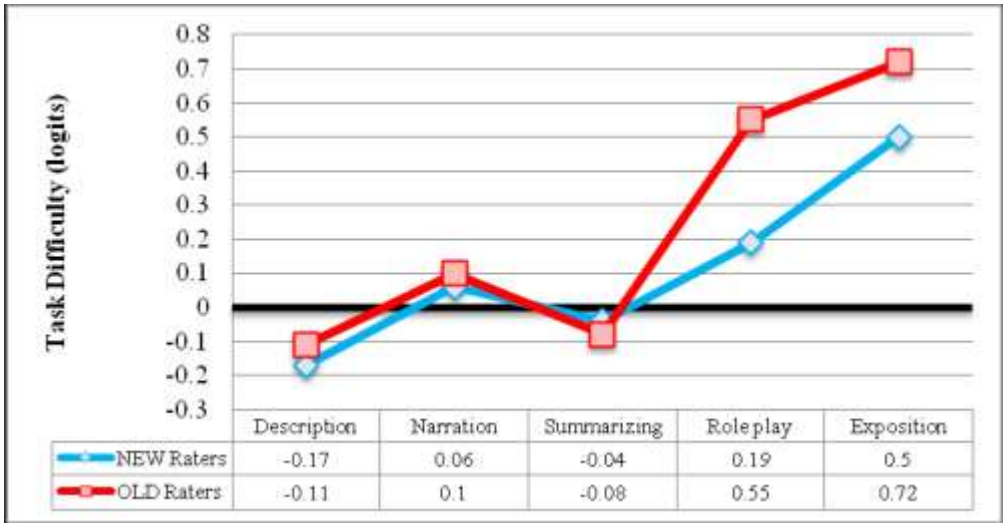
Task Difficulty and Bias Analysis Measure between Rater Groups and Tasks (Post-training)

Tasks	Raw score average	Fair average	Difficulty rating measure (logits) (Both groups)	Bias logits Z -Score		SE	Infit mean square	Outfit mean square	*r*
				NEW	OLD				
Description	2.78	2.72	-0.14	-0.14	-0.56	0.04	0.8	0.9	0.92
Narration	2.41	2.43	0.08	0.06	0.34	0.04	1.1	1.1	0.88
Summarizing	2.69	2.66	-0.06	-0.09	-0.21	0.05	1.1	1.1	0.93
Role play	2.03	2.15	0.37	0.72	1.14	0.05	1.2	1.2	0.88
Exposition	1.88	2.03	0.61	1.19	1.87	0.06	1.3	1.3	0.86
Mean	2.35	2.39	0.17	0.34	0.51	0.04	1.10	1.12	0.89
SD	0.39	0.30	0.31	0.58	0.99	0.00	0.18	0.14	0.03

Fixed (all same) chi-square: 229.52, $df = 4$, $p<0.00$
Task difficulty separation index: 0.75
Reliability index: 0.84

Figure 3 demonstrates task difficulty derived from NEW and OLD rater groups. As can be seen in the figure, the ratings of OLD raters were slightly with a little more fluctuation across tasks than NEW raters. The logit task difficulty for NEW raters measured ranging from -0.17 logits (description) to 0.50 logits (exposition) and a whole logit spread of 0.67 logits. For OLD raters, the range of task difficulty measure was rather a bit with more fluctuation than NEW raters, however a little wider in distance from 0. The task logit range was from -0.11 logits (description) to 0.72 logits (exposition) and a whole logit spread of 0.83 logits. This finding displayed that there were wider difficulty fluctuations with respect to scoring the oral tasks for OLD raters as compared to NEW raters. However, unlike the above-mentioned difference, similar to the pre-training phase, the figure also displays that both groups demonstrated relatively similar patterns in task difficulty measures. The description task was given the lowest difficulty measure and similarly the exposition task was given the highest difficulty measure by both rater groups.

A bias analysis, *columns five* and *six (Bias logits)*, was also run to investigate the interaction between rater groups and tasks. Unlike the pre-training phase, the analysis of the findings confirmed no trace of significant bias between the tasks and rater groups at the post-training phase. However, OLD raters, although were still within the acceptable range of biasedness, they were somehow very close to the borderline of biasedness in exposition (severity logit measure = 1.87). This is in a way that the same task severity was measured much less for NEW raters (severity logit measure = 1.19). At this phase, the least degree of biasedness for NEW raters was in description (logit value = -0.14), whereas for OLD raters again in description, (logit value = -0.56). This finding was rather confirmed by the outcomes of task difficulty measures in which exposition task was again identified the most difficult one. Figure 4 displays the bias analysis of the interaction between rater groups and tasks at the post-training phase.



	Description	Narration	Summarizing	Role play	Exposition
NEW Raters	-0.17	0.06	-0.04	0.19	0.5
OLD Raters	-0.11	0.1	-0.08	0.55	0.72

Figure 3. Task Difficulty Measures by NEW and OLD Rater Groups (Post-training)

A bias analysis, *columns five* and *six (Bias logits)*, was also run to investigate the interaction between rater groups and tasks. Unlike the pre-training phase, the analysis of the findings confirmed no trace of significant bias between the tasks and rater groups at the post-training phase. However, OLD raters, although were still within the acceptable range of biasedness, they were somehow very close to the borderline of biasedness in exposition (severity logit measure = 1.87). This is in a way that the same task severity was measured much less for NEW raters (severity logit measure = 1.19). At this phase, the least degree of biasedness for NEW raters was in description (logit value = -0.14), whereas for OLD raters again in description, (logit value = -0.56). This finding was rather confirmed by the outcomes of task difficulty measures in which exposition task was again identified the most difficult one. Figure 4 displays the bias analysis of the interaction between rater groups and tasks at the post-training phase.

In summary, some bias was shown for all the tasks of the study by both groups of raters. It should be noted that no significant bias interaction was observed at this phase. Quite drastically different from the pre-training phase, the facility of all oral tasks (description; narration; summarizing; role-play and exposition) for NEW raters was much more than those of OLD ones. The change of facility indices of tasks to score for NEW raters compared to OLD raters confirmed the effectiveness of the training program in familiarizing the raters of the oral tasks. In particular, NEW raters seemed to have benefited more from the training program than OLD raters due to their higher readiness, willingness, and attention to the instructed principles of the training program. A hypothetical reason could be that OLD raters had less tendency to accept further education from authorities due to their over

self-confidence. Thus, that is why NEW raters got more out of the training program than OLD raters. This is something which has also been reflected in their verbal protocol productions, that is, they reiterated that they were either not willing enough or rather skeptical in adapting their rating approach to that of the trainer thus changing the way they did their ratings.

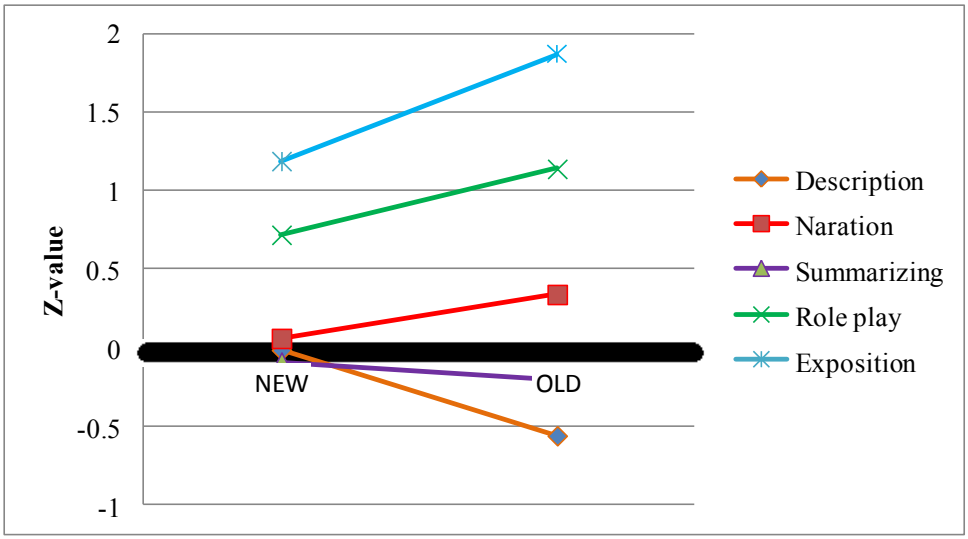


Figure 4. Rater-task Bias Interaction (Post-training)

3. Does test takers' score variability reflect their true speaking ability?

In order to identify whether the test takers' score variability is more dependent to the raters' scoring, the tasks in use or any other variables, an ANOVA, using the results of the FACETS analysis including raters severity in scoring (in logits) and oral tasks in difficulty (in logits), was run to determine whether test takers' score variance is due to the raters' scoring, task difficulty or other variables (Table 8).

The outcome of the table showed that both raters and oral tasks had significant contribution to test takers' oral score variability. However, a careful look at the residual on the last line of the ANOVA table implied that there is some amount of variance in test takers' score that could be related to either raters or oral tasks. Therefore, test takers own oral ability, as the high amount of residual showed, also had a determining role in their score variance. The higher amount of obtained residual, as compared to the effect of raters and tasks, demonstrated that test takers ability acted as a more significant role in test takers oral ability rather than other involving factors.

Table 8

ANOVA Table of Factors Influencing Test Takers' Task Performance

	Sum of Squares	*df*	Mean Square	F	Sig.
Rater	8.489	19	2.122	1.512	0.000
Task	12.601	5	6.331	2.244	0.000
Rater * task	5.312	95	1.68	0.94	Not Sig.
Residual	19.891				

In summary, the outcome of this part of data analysis demonstrated that a second language oral test represents variability across various test tasks. The participants in this study performed differently across all tasks except narration-role play (pre-training phase), and description-summary and role play-exposition (post-training phase). This variability might be attributed to the different requirements of each task both cognitively and linguistically which influenced their performance. For instance, for role play, a dialogic task, the rater was there to interact with the test takers; however, for the rest of the tasks, being monologic in nature, the test takers were constrained just by a set of pictures, figures, etc. without having access to any linguistics support.

4. Is there any significant relationship between task difficulty and raters' interrater reliability in scoring?

In another data analysis, in order to investigate whether task difficulty affects test reliability or not, the reliability of raters' scoring of each task was measured. The outcome of data analysis will reveal whether there is a relationship between task difficulty and rating reliability of each task. In this respect, interclass correlation coefficient was run using the data obtained at the post training phase to measure interrater reliability among raters in terms of scoring various difficulty tasks. Table 9 displays the interclass correlation coefficient representing interrater reliability among raters' scoring in each task type.

Table 9

Interclass Correlation among Raters' Scoring and Task Difficulty for Each Task

Task	Description	Narration	Summarizing	Role Play	Exposition
Difficulty (logits)	-0.14	0.08	-0.06	0.37	0.61
Reliability	0.75	0.69	0.71	0.66	0.64

Moreover, in order to make sure whether there is a significant difference between the obtained reliability measures, an ANOVA was run (Steiger, 1980) to investigate any possible significant difference. Table 10

displays the ANOVA outcome investigating the significant difference among interrater reliability measures for each task.

Table 10

ANOVA Table Investigating the Interrater Reliability Measures Related to each Task

	Sum of Squares	*df*	Mean Square	F	Sig.
Between Groups	0.001	4	0.000	0.768	0.549
Within Groups	0.042	95	0.000		
Total	0.044	99			

4.2. Discussion

The outcome of the table showed that there is no difference in terms of reliability measures between the tasks used in the study. This indicates that task classification on account of their difficulty levels has no effect on the amount of rater consistency in scoring. In other words, task difficulty has no significant impact on interrater consistency and agreement.

The outcome of the first and second research questions dealing with raters' biases to the tasks of various levels of difficulty indicated significant differences between NEW and OLD raters in their biases to the oral tasks at the pre-training phase. In this respect description task was identified to be the easiest and the exposition the most difficult. Such finding is fairly consistent with the one found by Trace, Janssen and Meier (2017) who claimed that raters' performances vary from task to task due to their difficulty measures. This finding is also similar to that of Skehan and Foster (1999) who found variation in terms of task difficulty between argumentation and instruction tasks. The finding is also in line with that of In'nami and Koizumi (2016) who found differences in task difficulty measures in the ACTFL oral test. At the pre-training phase, NEW raters were shown to have significant severity to exposition task, whereas OLD raters were at the optimum range of biasedness-although at the borderline. Both New and OLD raters demonstrated the highest severity towards exposition task confirming the previous finding in which exposition was shown to be the most difficult task. OLD raters were shown to demonstrate higher severity in most of the tasks compared to NEW raters.

Such difference in scoring tasks between NEW and OLD raters at the post-training phase was once again shown to be significant. Similarly, the description task was shown to be the easiest task, whereas the exposition task was the most difficult one. Nevertheless, unlike the pre-training phase, data analysis showed no significant difference between NEW and OLD raters' biases in scoring tasks with respect to various difficulty measures. The findings showed that there were wider difficulty fluctuations with respect to scoring the oral tasks for OLD raters as compared to NEW ones. It should be

noted that no significant bias interaction was observed after training. The facility of all oral tasks for NEW raters was much more than those of OLD ones. The change of facility indices of tasks to score for NEW raters compared to OLD raters confirms the effectiveness of the training program in familiarizing the raters of the oral tasks. Such finding is closely in line with that of Khabbazbashi (2017) who found the constructive effectiveness of training in reducing raters' variability in scoring. This indicates that NEW raters seem to have benefited more from the training program than OLD raters due to their higher readiness, willingness and attention to the instructed principles of the training program. A hypothetical reason could be that OLD raters had less tendency to accept further education from authorities due to their over self-confidence. Thus, that is why NEW raters got more out of the training program than OLD raters. The more constructive impact of the training program for NEW raters compared to OLD ones confirms the research finding by Attali (2016) and Bijani (2010) who, in separate studies, found that inexperienced raters benefited more from training than experienced ones.

The variety of test tasks used in the study caused the rater groups display various rating behaviors. Not only did the raters display different severity measures, but also they adopted different evaluating criteria in different tasks. This finding shows that the use of various test tasks can be effective in eliciting various rating behaviors. The outcomes of this study also demonstrated that the raters were more lenient in scoring description thus the test takers received higher scores in description than the other tasks specifically compared to exposition and narration. Several possible explanations can be suggested for the test takers' better performance on description compared to other tasks. One possibility could be that performing a task which requires answering questions is more common in interview and speaking tests and most typically students are already familiar with this kind of speaking task than narrating a story of sequential pictures or dealing with figures and diagrams. This finding is in line with that of In'nami and Koizumi (2016) who found a higher fluency for the students dealing with oral interview than other tasks. A second possibility could be that description tasks are more structured and that's why students are capable of generating more fluent speech. This hypothetical reason is consistent with that of Skehan and Foster (1999) who argued that those tasks which contain a more organized structure will result in a more fluent performance.

The outcome of the third research question which was attributed to the effectiveness of various factors in test takers' score variability, the results showed that although raters and tasks have significant contribution to test takers' score variance, the effect of test takers' own oral ability is a much more determining factor among the three. The high amount of obtained

residual, as compared to the effect of raters and tasks, demonstrates that test takers ability acts as a more significant role in test takers oral ability rather than other involving factors. This indicates that test takers' score variation is more influenced by their own performance ability than the nature of tasks or raters' biases. Although there is paucity of research in this respect, the outcome is consistent with that of Winke, Gass and Myford (2012) who found the impact of test takers more significant in their score variation than other intervening variables. This finding is also consistent with that of May (2009) who found similar outcome about the relationship between task type differences and significant task difficulty; however, this effectiveness of task type differences on test-takers' scores was found to be little effective. Similar finding was also obtained by Khabbazbashi (2017) who found a close relationship between test takers' ability and task difficulty types. The outcome of the study regarding the relationship between task difficulty and magnitude of interrater reliability among raters, displayed no relationship between the two which shows that task difficulty does not affect interrater reliability among raters. This outcome is also reflected in a study by Ling, Mollaun and Xi (2014) which found no significant interaction between language type and task difficulty.

The outcome of the fourth research question, dealing with the relationship between task difficulty measures and raters' interrater reliability, showed no significant difference in terms of reliability measures between the tasks used in the study. This indicates that task classification on account of their difficulty levels has no effect on the amount of rater consistency in scoring. In other words, task difficulty has no significant impact on interrater consistency and agreement. This finding contradicts with that of Ahmadian and Tavakoli (2011) who found that as task difficulty increases, raters tend to display lower measures of interrater reliability.

5. Conclusion and Implications

The findings of this study on the basis of statistical MFRM outcomes demonstrated the usefulness of this analytical approach in detecting rater effects and demonstrating the consistency and variability in rater behavior aiming to evaluate the quality of rating. MFRM can provide raters with rapid feedback on their instability and thus to apply adjustments on raters' behaviors based on that feedback. This study showed that rating oral proficiency tasks is context-specific. The analysis confirmed that the nature of second language oral construct is not constant, thus different results are achieved using different oral tasks. Test difficulty identification is complex, difficult and at the same time multidimensional (Lumely & McNamara, 1995). However, test takers' perceptions could be considered as a reliable factor for determining task difficulty.

This study showed that tasks are different with respect to their difficulty measures. Besides, various groups of raters have biases to different tasks in use. Consequently, training programs can reduce raters' biases and increase their consistency measures. The findings also showed that test takers' performance ability is the foremost significant factor in determining their score variation as compared to other intervening variables (e.g., raters or oral tasks); therefore, the remaining intervening variables can be modified and reduced by establishing effective training programs. Additionally, task difficulty measures were shown not to have any impact on measures of raters' interrater reliability. This suggests that decisions makers had better not be concerned about kinds of tasks in use for the sake of achieving acceptable measures of reliability in assessment.

The outcomes suggest that decision makers had better not be concerned about raters' expertise. In other words, although decision makers commonly use experienced raters for the sake of achieving higher measures of reliability in assessment, the outcome of the study showed that there is no significant difference between experienced and inexperienced raters after training and even inexperienced raters showed less bias and higher consistency measures in assessment. Through rater training programs, rater effects and variability can be controlled. Thus decision makers had better establish rater training programs to increase rater consistency and reduce their biases in measurement.

However, this finding also must not be misinterpreted as a key factor of establishing a hierarchical order of task difficulty solely on the basis of test taker's testing intuition. Besides, generalizations must be done with great caution. It is important for performance assessment test to take into consideration the effect of task characteristics and most importantly, performance conditions in estimating the performance ability of test takers. This research got benefit from five oral tasks in the direct and indirect version respectively. The replication of the research adopting the use of other types of oral tasks could be done in future studies. Besides, it considered six factors which were hypothesized to be influential in task difficulty dimension measures, further studies could be run investigating other task testing dimensions on their possible effectiveness of task difficulty.

References

Ahamadian, M.J. & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research, 15*(1), 35-59.

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

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

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.

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

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*(4), 347-368.

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

Gardner, R. (1992). *Task classification, methodology and task selection.* Unpublished manuscript, Department of Linguistics and Applied Linguistics, University of Melbourne.

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.

Jeong, H., & Hashizume, H. (2011). Testing second language oral proficiency in direct and semidirect settings: A social-cognitive neuroscience perspective. *Language Learning, 61*(3), 675-699.

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing, 34*(1), 23-48.

Kuiken, F., & Vedder, I. (2014). Raters' decisions, rating procedures and rating scales. *Language Testing, 31*(3), 279-284.

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.

Leaper, D. A., & Riazi, M. (2014). The influence of prompt on group oral tests. *Language Testing, 31*(2), 177-204.

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*(4), 479-499.

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

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

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

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair task performance. *Language Testing, 19*(3), 277-295.

Robinson, P. (2001). Task complexity, task difficulty and task production: Exploring interactions in a componential framework. *Applied Linguistics, 21*(1), 27-57.

Skehan, P (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*(1), 93-120.

Steiger, J. H., (1980). Test for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245-251.

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

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.

Bibliographic information of this paper for citing:

Bijani, H. (2018). Effectiveness of a face-to-face training program on oral performance assessment: The analysis of tasks using the multifaceted Rasch analysis. *Journal of Modern Research in English Language Studies, 5*(4), 27-53.