Volume 2, Issue 2, Autumn-Winter 2018 (Biannual – Serial No. 4)

The Investigation of Rater Expertise in Oral Language Proficiency Assessment: A Multifaceted Rasch Analysis[1]

Houman Bijani*[2]

Received: 2019-05-13 | Revised: 2019-07-16 | Accepted: 2019-08-27

Abstract

Since scoring oral language proficiency is performed by raters, they are an essential part of performance assessment. One important feature of raters is their teaching and rating experience which has attracted considerable attention. In a majority of previous studies on rater training, extremely severe or lenient raters, benefited more from training programs and thus results of this training showed significant severity/leniency reduction in their rating behavior. However, they mostly investigated the application of FACETS on only one or two facets and few have used a pre, post-training design. Besides, empirical studies have reported contrasting outcomes, not showing clearly which group of raters does rating more reliably than the other. In this study, 20 experienced and inexperienced raters rated the oral performances produced by 200 test-takers before and after a training program. The results indicated that training leads to higher measures of interrater consistency and reduces measures of biases towards using rating scale categories. Moreover, since it is almost impossible to completely eradicate rater variability even if training is ap-

[1] DOI: **10.22051/lghor.2019.26072.1123**

[2] Assistant Professor, Department of English Language Teaching, Zanjan Branch, Islamic Azad University, Zanjan, Iran, (Corresponding Author); houman.bijani@iauz.ac.ir

plied, rater training procedure had better had better be regarded as a procedure to make raters more self-consistent (intrarater reliability) rather than consistent with each other (interrater reliability). The findings of this study indicated that inexperienced and experienced raters' rating quality improved after training; however, inexperienced raters underwent higher consistency and less bias. Hence, there is no evidence that inexperienced raters should be excluded from rating solely because of their lack of adequate experience. Moreover, Inexperienced raters, being more economical than the experienced ones, cost less for decision-makers for rating. Therefore, instead of charging a bulky budget on experienced raters, decision-makers had better use the budget for establishing better training programs.

Keywords: Bias; Interrater consistency; Intrarater consistency; Multi-faceted Rasch measurement (MFRM); Rater expertise

Introduction

In scoring second language speaking performance, rater variability has been identified as a potential source of measurement error which might interfere with the measurement of test-takers' true speaking ability (McNamara, 1996). Therefore, rater effects are required to be taken into consideration in order to measure test-takers' speaking ability appropriately. One important, related rater feature that has been demonstrated to influence test-takers' test scores is rater background. Various groups of raters may differ in the judgment of learners' second language ability depending on their background and the criteria they apply (Barrett, 2001). From among all rater effects, oral language teaching and rating experience are the variables which have attracted the most attention. One of the most critical worrisome factors in raters' scoring is whether they have been adequately trained or have had enough expertise in assigning accurate scores (Winke, Gass & Myford, 2012). According to Cumming (1990), experience refers either to the period of time the rater has been rating or to the amount of rating the rater has performed, whereas expertise refers to the raters whose ratings are consistently good. Although experiences and expertise are related issues, they are different in a way that experience may or may not lead to expertise.

Although some inexperienced raters may represent acceptable rating patterns even before or obviously after training, those who have extremely severe or lenient scoring patterns or even inconsistent ones seem to become more similar to the experienced ones after a few rating sessions. Studies have reported a gradual but steady increase in rater consistency over time as inexperienced raters familiarize themselves with the scoring system (Bonk & Ockey, 2003; Lim, 2011). This process continues until a certain amount of variability in raters' consistency and severity remains regardless of the experience. That is, even experienced raters do not perfectly correlate with each other (Bijani, 2010; Kim, 2011). Therefore, it could be concluded that further help, in addition to experience, can benefit raters with inconsistent and biased scoring patterns; however, the influence is limited.

Literature Review

Rater Expertise and Oral Assessment

During the last 15 years, researchers have shifted their focus of attention on features of raters which may influence their ratings (e.g., Barrett, 2001; Bonk & Ockey, 2003; Caban, 2003). Rating experience is the variable which has attracted the most attention in assigning accurate scores to test-takers' oral performances (Winke et al., 2012). This depends on the experiences that a rater has had, cognitive factors, the characteristics of the rating criteria, and the rating environment.

A variety of studies on experienced and inexperienced raters' performances have indicated higher inter-rater consistency following training (Ahmadi & Sadeghi, 2016; Attali, 2016; Bijani & Fahim, 2011; Cumming, 1990). Commonly, in all these studies, extremely severe or lenient inexperienced raters benefited from the training program and thus modified their rating behavior. In a study by Bijani (2010) on the effect of rater training on raters' inconsistency in scoring test-takers' written language proficiency, the consistency of inexperienced raters improved much more after training compared to experienced raters. Some studies have found that inexperienced and experienced raters use different rating approaches to evaluate students' performances. For example, Kim (2015) found that experienced trained raters use an approach commonly known as the Funnel Model (a process in which raters score all performances on one feature and then categorize them on the basis of other features) to guide their judgments.

Difference between Experienced and Inexperienced Raters

Several studies have found differences between inexperienced and experienced raters in their scorings and the use of rating strategies (Cumming, 1990; Davis, 2009; Huang, Huang, & Hong, 2016; Leaper & Riazi, 2014; Nakatsuhara, 2011). Huang et al. (2016) compared the ratings of trained and untrained raters from two backgrounds: experienced English teachers and non-teachers. They found that training was a more significant variable than background in terms of reliability. However, they did not report any differences with regard to the overall differences between groups in rater severity. Nakatsuhara (2011) compared the holistic ratings of novice and expert teachers on 114 test-takers' oral language samples, and found that novice teachers' ratings were lower than expert teachers'. On the other hand, she found that rater trainers were severer in their ratings than inexperienced raters. Similar outcomes were reported by Leaper and Riazi (2014) who found that novice raters were significantly more lenient in their ratings of coherence and fluency, and by Davis (2009) who found that experienced raters were significantly severer in their ratings than inexperienced raters in holistic scores of the speaking ability (ACTFL).

Ahmadi and Sadeghi (2016) studied a group of experienced and inexperienced language teachers and provided both with one-day training in rating oral performance tests. Using a multifaceted Rasch measurement (MFRM), they found that inexperienced raters' ratings were relatively severer than those of the experienced ones' with respect to politeness and pronunciation, and that they overfitted the given model, i.e., there was insufficient variability in their ratings. However, experienced raters were likely to have more diversity in their ratings. They concluded that there exist factors, other than the ones in the rating scale, which affect the raters' scoring. In their study, In'nami and Koizumi (2016) found that experienced teachers have employed certain rating criteria which are different from those of inexperienced teachers. They argued that as, inexperienced teachers focused more on content, they mainly focused on grammatical and pronunciation errors. However, a contrasting finding was observed by Davis (2016) who found out that experienced teachers focused on communicative assessment, whereas inexperienced teachers emphasized grammar and pronunciation. These findings suggest that raters' background may play a role in how they perform rating. Galloway (as cited in Bonk & Ockey, 2003) investigated the ratings of experienced and inexperienced Spanish raters scoring 10 students' response to speech sample questions. The results demonstrated that inexperienced raters were more lenient than experienced ones. In contrast, Lim (2011), in a similar study, found that inexperienced raters were severer than experienced ones. Kyle, Crossley, and McNamara (2016) studied the inter-rater reliability of four groups of raters, professional and lay raters both with and without training. It was shown that trained raters enjoyed more inter-rater reliability than untrained ones regardless of their background and level of expertise.

Attali (2016) studied four inexperienced raters scoring compositions both before and after the rater training program. The results demonstrated the clarification of scoring criteria, modification of their awareness, and awareness of the need for inter-rater agreement which consequently brought inexperienced raters more in line with the experienced ones. Caban (2003) used MFRM in a research study on group oral assessment for Japanese EFL learners. He used a group of expert and non-expert raters in assessing the test-takers' spoken language. The results demonstrated a high variability among raters. It was also revealed that raters tend to become severer with experience. Van Moere (2012) compared the performance of English experienced and inexperienced raters scoring Chinese students studying English in the US. The findings showed no significant difference between the two groups of raters with regard to the degree of severity. Barkaoui (2011) classified raters into three groups of expertise. With the help of verbal protocols, they found qualitative differences in the way they rated. Proficient raters had fewer interruptions and could make their judgments after they finished the work. They also produced more comments and the scores they awarded better matched the scoring rubric. Therefore, although all the raters in the study were experienced ones, there still existed some qualitative differences in their rating approaches which were attributed to their expertise.

A number of research studies have shown that rater training minimizes rater effects (e.g., Attali, 2016; Bijani, 2010; Davis, 2016; In'nami & Koizumi, 2016). Davis (2016) found that the provision of feedback on raters' scoring behaviors could assist them in becoming more consistent on the subsequent ratings. Rater training can help raters better understand the categories and criteria of rating scales which might influence their rating behavior (Kuiken & Vedder, 2014). Accordingly, in the absence of rater training programs, raters with various levels of expertise may assign different scores to the language being tested (Bijani & Fahim, 2011; Cumming, 1990), whereas extended training programs will help them develop a common reference framework.

MFRM introduced by Linacre (1989) 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 (Khabbazbashi, 2017). In this approach, rater variation is seen as an inevitable part of the rating process, but not an obstacle to measurement. Proponents of the Rasch approach to measurement claim that raters cannot be trained to achieve similar levels of severity since estimates of test-taker ability are said to be independent of the severity of the particular raters who happen to rate those particular test takers (Wright & Linacre, 1994). However, as Khabbazbashi (2017) argues, the differences between raters can be found out with respect to severity and random error; therefore, training is recommended for raters who are identified as misfitting by the Rasch analysis to make raters more self-consistent (intra-rater consistency) rather than aiming for interrater consistency.

However, most of the studies conducted so far have investigated the application of FACETS on only one or two facets, including the study of rater's severity/leniency on specific test-takers (Barkaoui, 2011), on task types (In'nami & Koizumi, 2016), and on certain rating time (Gan, 2010). Thus, no study so far has included the facets of test-takers' ability, raters' difficulty, group expertise, and scale criterion category all in a single study along with their bilateral effects. Besides, while a few studies have looked at the differences between trained and untrained raters in speaking assessment (Khabbazbashi, 2017; Kim, 2011; Gan, 2010) and other contexts (Ahmadi & Sadeghi, 2016), few studies have used a pre- and post-training design.

Empirical studies on the effects of training and experience have reported contrasting outcomes (Lim, 2011; Winke et al., 2012), consequently, a better understanding of how training and experience could be mixed to add more reliability to scoring seems essential. While there are some general differences between experienced and inexperienced raters (e.g., Attali, 2016; Bijani, 2010), there is little research dealing with this issue and that which group of raters does the rating job more reliably than the other. This study aims to resolve the above-mentioned shortcomings by taking a meticulous analytical approach investigating the four mentioned facets using a pre, post-training design to investigate the change in the behavior of experienced and inexperienced raters. Therefore, the following research question was formulated formed:

RQ: Is there any significant difference between experienced and inexperienced raters in terms of severity, consistency, and bias measures before and after training?

Method

Participants

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

Twenty Iranian EFL teachers, including 10 males and 10 females, ranging in age from 24 to 58 years participated as raters. In order to fulfill the requirements of this study, the raters had to be classified into two groups of experienced raters and inexperienced ones to investigate the similarities and differences among them and the likelihood of 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 five years of experience in teaching and passed less than the four core courses related to the 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 five years of experience in teaching and passed all the four core courses plus at least two selective courses related to ELT major. Hereinafter we call these raters as Old.

Instruments

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.

Scoring rubric

Each test taker's task performance was assessed using the Educational Testing Service (2001) analytic rating scale. In Educational Testing Service (2001) scoring rubric, individual tasks are assessed using appropriate criteria including *fluency*, *grammar*, *vocabulary*, *intelligibility*, *cohesion* and *comprehension*.

Procedure

Pre-training Phase

Prior to collecting 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 (pre, post-training). All the raters participating in this study were given one week to submit their scorings.

Rater Training

After the pre-training scoring 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. In addition to the norming 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 significant bias and were reminded individually to mind the issue. 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, with the raters with an infit mean square value below 0.6 being too consistent (overfit the model) and those with an infit mean square value of above 1.4 being inconsistent (underfit the model). Therefore, the raters were pointed out individually on the issue if they were identified as misfitting.

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, data were elicited from the second half of the test-takers (including 100 students).

Data Analysis

In order to investigate the research questions, a pre-, post-test 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 using MFRM during two scoring sessions for the four test facets including test-takers, rater, rater group, and rating criterion, and their interactions to investigate variations in rater behavior and rater biasedness. The scoring patterns of the two groups of raters (inexperienced and 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 test takers was investigated to identify any hypothetical difference of the impact of factors on test takers' oral performance scores.

Result

The rater group bias analysis was measured for each expertise group. Table 1 displays the raters' measurement report for New and Old raters along with the chi-square results and the significance level at the pre-training phase.

Column one *(Rater groups)* demonstrates the rater groups. Column two *(Observed average score)* represents the raters' mean scores of each group of expertise given to the test-takers. The statistics shows that New raters assigned higher scores than Old raters (New: 23.58 vs. Old: 21.21), thus indicating that they were more lenient than Old ones.

Column three *(Fair average)* demonstrates the extent to which the mean ratings of raters' scores in each group differ. For instance, here, the mean rating of Old raters is 21.21 and the fair average is 21.32. Similarly, the mean rating of New raters is 23.58 and the fair average is 23.01. The data show that the two groups' scorings are 2.37 raw scores apart when comparing the mean ratings and 1.69 raw scores when comparing their fair averages.

Column four *(Obs-Exp score)* displays the total observed score for all the 100 test takers at the pre-training phase rated by the raters in each expertise group minus the total expected score for the test takers rated by the same group.

Column five *(Bias logit)* was run to study group severity/leniency measures. The result revealed that Old raters were severe in their ratings (logit = 0.44), whereas New raters were rather lenient (logit = -0.68). In general, the bias measure for New raters was more than that of the Old ones. The mean bias value (in logits) was -0.12; thus, the rater groups displaying more than half a logit value above or below the mean logit (between -0.62 and 0.38) would be considered as either too severe or too lenient (Wright & Linacre, 1994). In this respect, Old raters were identified as too severe and New raters as too lenient.

Column six *(SE)* displays the standard error of bias estimation. This value for Old and New raters is 0.04 and 0.03, respectively. The small value of SE provides evidence for the high precision of measurement.

Columns seven and nine *(Infit and outfit mean square)* display the 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 is simply used to determine within-rater consistency *(Intra-rater consistency)* which indicates the extent to which each rater ranks the test takers consistently with his/her true ability. Fit statistics is categorized into two subparts entitled *infit* and *outfit* statistics. 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. In other words, it is the average 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 (Eckes, 2015). Fit statistics has the expected value of 1 and a range of zero to infinity; however, there is no straightforward rule or an absolute or universally definite range for interpreting fit statistics value or for setting the upper and lower limits; therefore, the acceptability of fit is examined on a judgmental basis not solely on a statistical one. The acceptable range of fit statistics, according to Wright and Linacre (1994), is within 0.6 to 1.4 logit values. Therefore, in order to investigate the raters' fit statistics value, the raters who are placed below this range are *overfit* or *too consistent*, and those above this range are *underfit* or *too inconsistent*.

The infit mean square was 1.3 for the New rater group and 0.8 for Old raters. This finding demonstrates that although both groups of raters, according to Eckes (2015), are at the acceptable fit statistics range, New raters are relatively on the borderline of inconsistency (Infit MnSq. = 1.3), whereas this statistics for Old raters is 0.8, showing that they have more tendency towards homogeneity and consistency in rating.

Also, columns eight and ten *(Z-scores)* display group rater bias estimate. *Bias* is the difference between expected and observed ratings of the obtained data which is then divided by its standard error to achieve the z-score (McNamara, 1996). The most preferable amount of z-value is 0 which indicates that the data match the expected model, thus there exists no bias on the side of raters. According to McNamara (1996), z-values between ±2 are considered as the acceptable range of biasedness; therefore, both groups of raters were considered as having significant biasedness but to opposite directions. i.e., Old raters had the tendency towards significant severity (Z_{Old}= 2.22) while New ones to significant leniency (Z_{New}= -2.64). The result also indicates that the amount of biasedness for New raters at the pre-training phase was more extreme than that for Old raters.

Moreover, in order to examine to what extent the two rater groups were similar to each other in ranking the test takers, a single rater-rest of the raters' correlation (point-biserial correlation) was run (column eleven). The results were quite different in a way that the correlation index for Old raters was 0.61, whereas it was 0.39 for New raters, showing that Old raters tended to rank order the test takers in a more precise and consistent way.

However, the logit severity estimates do not themselves tell us whether the differences in severity/leniency estimates are meaningful or not; consequently, 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. 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 severity among the raters. 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 (Eckes, 2015). In case the raters are equally severe, the standard deviation of the rater severity estimates should be equal to or smaller than the mean estimation error of the entire data set which results in a separation index of 1.00 or even less. In this study, the *separation index* of 3.66 for New raters and 3.49 for Old ones demonstrated that the rater groups could be classified into nearly three and a half groups of severity measures. The *reliability index* of 0.90 for both groups of expertise, which is quite high, provides further proof for the precision of this separation that the raters were reliability separated into various levels of severity measures. The fixed chi-square value for all the 20 raters rating the test takers' oral performance was measured at the pre-training phase as well ($X^2_{(19, N=20)} = 46.05$, $p < 0.00$). The finding suggested that the rater groups differ significantly and are not at the same level of severity.

Table 1.
Old and New Rater Group Measurement Report (Pre-Training)

Rater group	Observed average score	Fair average	Obs-Exp score (logit)	Bias Logit	SE	Fit statistics				Correlation Point biserial
						Infit MnSq.	Z score	Outfit MnSq.	Z score	
Old raters	21.21	21.32	-3.55	0.44	0.04	0.80	2.22	0.70	2.14	0.61
New raters	23.58	23.01	3.84	-0.68	0.03	1.30	-2.64	1.30	-2.46	0.39
Mean SD	22.39 1.67	22.16 1.19	0.14 5.22	-0.12 0.79	0.03 0.00	1.05 0.35	-0.21 3.43	1.00 0.42	-0.16 3.25	0.50 0.15

Old: Experienced raters New: Inexperienced raters
Fixed (all same) Chi-square: 46.05, *df* = 1, *p* < 0.00
Separation index (Old): 3.49 Reliability (Old): 0.90
Separation index (New): 3.66 Reliability (New): 0.90

In order to make sure whether the difference in overall severity between the two groups of rater expertise is significant and to determine whether they ranked test takers in the same way, a Mann-Whitney U test was run. Table 2 displays the Mann-Whitney U test results for New and Old raters at the pre-training phase.

Table 2.
Mann-Whitney U Test for New and Old Raters' Severity Index (Pre-Training)

Rater group	N	Σ Rank	Mean Rank	Std. D.	Separation	Reliability
New	10	128	12.8	0.82	4.63	0.91
Old	10	82	8.20	0.61	2.69	0.88
Z: -2.09, $p < 0.05$						

The less the raters' ranks are, the more severe they will be (column four). That is, Old raters (Mean Rank = 8.2) were identified to be severer than New ones (Mean Rank = 12.8). As the table shows, Old raters are severer than New raters (Z= -2.09) which is also significant at $p < 0.05$ showing a significant difference between New and Old raters in terms of severity. This severity difference between New and Old raters represents a large separation index. The range of severity estimate is much larger for New raters than Old raters by examining the standard deviation of the severity estimates which is 0.82 for New raters and 0.61 for Old ones. This variability is also reflected in the separation indices for the two groups, i.e., 4.63 for New raters and 2.69 for Old raters. The higher severity of Old raters increases the overall rater severity spread of all 20 raters. The reliability index (column seven) demonstrates the reliability measure of the separation index. This measure is fairly high for both rater groups.

In order to have a better picture of the systematic pattern of raters and test taker bias interactions, a rater-test taker bias interaction analysis for various ranges of bias logit was performed. Table 3 displays the bias between raters and test-takers for various logit range values. At the pre-training phase, New raters had a more significant bias towards test-takers than Old ones (337 to 307). This shows that Old raters were less biased towards test-takers than New ones. This can be due to the fact that Old raters used better strategies to judge the test-takers in accordance with their true oral ability. The mean number of significant bias interactions for New raters was found to be 33.7 and for Old ones to be 30.7.

Over half of the interactions, 361, occurred around the mean, i.e., from 0.99 to 0.99 logit values. This can be on account of the fact that a majority of rater-test-taker interactions were clustered in that range. The table displayed that raters were likely to demonstrate more biasedness towards higher-ability test-takers than the lower-ability ones. There were 332 bias interactions above 0.00 and 312 bias interactions below 0.00. Bias interactions for higher-ability test-takers were more likely be severe than lenient (186 severe and 146 lenient); similarly, bias interactions for lower-ability test-takers were more likely to be lenient than severe (171 lenient and 141 severe). The same pattern is applicable even at the extreme ends of the scale (logit values from -3.0 to -3.99 and from 3.0 to 3.99) as well. The highest-ability test-takers received 12 out of 18 severe interactions, whereas the lowest-ability test-takers attracted 9 out of 13 lenient interactions. The reason for this interaction tendency is not quite clear; however, it might be due to raters' increasing expectations of test-takers as the ability of the test-takers becomes higher, thus making their judgments severer. For lower-ability test-takers, perhaps raters would benefit test-takers to compensate for their lack of proficiency. This finding is relatively in line with that of Kondo-Brown (2002) who found a similar pattern of rater-test-taker biasedness, but in writing.

Table 3.
Bias Interaction Frequency between Raters at Each Expertise Level and Test-takers for Various Bias Logit Range (Pre-Training)

Logit band ➔ Rater ⬇	-3.0 to -3.99		-2.0 to -2.99		-1.0 to -1.99		0.0 to -0.99		0.0 to 0.99		1.0 to 1.99		2.0 to 2.99		3.0 to 3.99		Total (Scale in whole)
Severity/Leniency	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	
New raters	1	5	7	19	18	21	39	54	38	58	23	18	14	14	3	5	337
Old raters	3	4	14	11	9	26	50	31	57	34	28	12	14	4	9	1	307

Mean rater-test-taker bias interaction for:
New raters: 33.7
Old raters: 30.7

In order to study the rater group biasedness to each particular category of the rating scale, a bias interaction analysis with regard to each category was performed. Table 4 demonstrates the frequency of each rater group interaction to each category of the rating scale. Since there were 10 raters in each expertise group and 100 test-takers at the pre-training phase as well as six scale categories, a 6000 interactional frequency was obtained. Upon a chi-square analysis to analyze the data (Steiger, 1980), the outcome revealed a significant difference between the two groups in terms of their biases to scale categories. Data analysis revealed that for more difficult scale categories (cohesion and intelligibility), there appeared very large differences with regard to the raters' biases produced by each group in a way that New raters showed a lot more biasedness than Old ones. This can be due to the fact that the more difficult the scale categories become, the more the raters, especially the ones with inadequate knowledge and expertise, become confused, thus making more biased decisions.

Regarding fluency and vocabulary, which are the third and the fifth difficult category to rate, as shown in the facets map of variables, New raters tended to show more bias than Old ones. Although still significant, the difference was not as big as that of the previous two categories. The reason for this finding could be attributed to the fact that Old raters tend to have a global or optimistic view about language pronunciation and do not worry about non-native features of language as long as they do not seriously impede the flow of communication. New raters seemed to over-emphasize pronunciation and thus were negatively biased in this respect. They considered any deviance from the native-like pronunciation as errors; thus, they tended to mark down the performance accordingly. New raters still showed more bias in the category of comprehension than Old ones. This seems to be due to the fact that Old raters had a more careful consideration of meaning than New raters who seemed to have misunderstood the issue and focused their attention more on the structural aspect of language output which is discussed below.

The only category in which a reverse behavior was observed was grammar in which Old raters demonstrated more biases than New ones. This might be attributed to the fact that Old raters were cautious towards the correct and exact use of grammar and structural rules which made them too concerned about the issue. The result, of course, would make them become more biased in rating this category.

Table 4.
Frequency of Each Rater Group Interaction to Each Category of the Rating Scale (Pre-training)

Category	Frequency of bias interactions		Direction of difference	Sig.
	Old	New		
Cohesion	357	647	Old < New	**
Intelligibility	247	491	Old < New	**
Fluency	187	212	Old < New	*
Comprehension	146	159	Old < New	Not Sig.
Vocabulary	97	161	Old < New	*
Grammar	177	134	Old > New	*

* *P*<0.05
***P*<0.01

To sum up, the FACETS analysis of the rater groups revealed statistically significant differences between them. There appeared to be considerable differences in consistency and severity among the raters in each group when judging test takers' oral performances.

Table 5 displays the bias analysis of the raters in both groups of expertise at the post-training phase. As shown in column five *(Bias logit)*, the outcome of the table revealed that Old raters were still severe in their ratings (logit = 0.30), whereas New raters were rather lenient (logit = -0.10). In general, the bias measure for New raters (in whole) almost did not show any sign of biasedness and they were nearly at no level of severity/leniency. The mean bias value (in logits) was 0.10; thus, rater groups displaying more than half a logit value above or below the mean logit value (between -0.40 and 0.60) would be considered as either too severe or too lenient (Wright & Linacre, 1994). In this respect, unlike the pre-training phase, both rater groups were identified within the acceptable range of severity/leniency. However, the outcomes obviously display that New raters had a less interactional effect than Old raters who were a lot severer than New raters.

Column six *(SE)* displays the standard error of bias estimation. This value for Old and New raters is 0.5. The small value of SE provides evidence for the high precision of measurement. This value is rather low in the table showing the high degree of precision.

Columns seven and nine *(Infit and outfit mean square)* display whether the rater groups (New and Old) were at the acceptable fit index or not. The infit mean square for New raters measured 1.00 and for Old raters it was 0.84. This finding demonstrates that New raters were at the *perfect* degree of consistency with each other after training and proved the true effectiveness of the training program in bringing New raters in terms of consistency with each other. Old raters, except for one rater (Old8) who was detected as overfitting (infit MnSq. = 0.5), were also within the acceptable fit statistics range, too; however, they were not identified to be as consistent as New raters were. New raters tended to become a lot more consistent which provided enough evidence to claim that New raters benefited more from the administration of the training program than Old ones.

Also, columns eight and ten *(Z-scores)* display group rater bias estimate at this phase. As mentioned previously, according to McNamara (1996), z-values between ±2 are considered as the acceptable range of biasedness; therefore, the table obviously displays that both groups of raters, after training, were placed in the acceptable range of biasedness. Nevertheless, it is noteworthy to indicate that New raters (Z_{New}= -0.26) had a less interactional effect than Old raters (Z_{Old} = 0.58). After all, both groups, similar to the pre-training phase, had an interactional tendency to opposite directions.

Besides, in order to examine to what extent the two rater groups were similar to each other in ranking the test-takers, the point-biserial correlation was run once again (column eleven). The results demonstrated that Old raters had a correlation index of 0.63 which was a little better than that of the pre-training phase. However, the same result for New raters measured 0.73 which shows a drastic change. New raters at the pre-training phase had much less correlational index than Old raters, whereas after training, they took the lead, having a higher degree of correlation than Old raters. This outcome undoubtedly implies that New raters benefited much more from feedback and the training program than Old raters and that it was more useful for New raters than the other group.

The *separation index* of 1.71 for New raters and 1.83 for Old ones demonstrated that the rater groups could be classified into nearly two groups of severity measures. The *reliability indices* of 0.67 for New raters and 0.86 for Old ones, which were much lower than that of the pre-training phase, showed that the separation of raters into various levels of severity was less distinguishable after training. In other words, at the post-training phase, since the raters acquired more consistency and had less degree of severity and leniency and biasedness, it was rather difficult to clearly separate the raters into various levels of severity measures. However, the magnitude of this reliability index was higher for Old raters, showing that they were more precisely separated into various levels of severity measures. This adds more evidence for the usefulness of the application of the training program in bringing consistency and reducing biasedness and severity among the raters. Besides, the rather low value of reliability in this phase of the study (r = 0.67 for New raters and r = 0.86 for Old ones) indicates that the analysis could separate the raters into different levels of severity with less precision – due to the establishment of more consistency among the raters. The *fixed chi-square* value for all the 20 raters rating the test-takers' oral performance was measured after training as well ($X^2{}_{(19,\ N=20)}$ = 1.19, p>0.05). The result suggested that the rater groups, at the post-training phase, were at the same level of severity.

Table 5.
Old and New Rater Group Measurement Report (Post-training)

Rater group	Observed average score	Fair average	Obs-Exp score (logit)	Bias Logit	SE	Fit statistics				Correlation Point biserial
						Infit MnSq.	Z score	Outfit MnSq.	Z score	
Old raters	21.39	20.74	21.94	0.30	0.05	0.84	0.58	0.88	0.59	0.63
New raters	23.07	22.37	22.88	-0.10	0.05	1.00	-0.26	1.03	-0.28	0.73
Mean SD	22.23 1.18	21.55 1.15	22.41 0.66	0.10 0.28	0.05 0.00	0.92 0.11	0.16 0.59	0.95 0.10	0.15 0.61	0.68 0.07

Old: Experienced raters   New: Inexperienced raters
Chi-square:1.19, *df*=1, *p*>0.05
Separation index (Old): 1.83   Reliability (Old): 0.86
Separation index (New): 1.71   Reliability (New): 0.67

Once again, in order to ensure whether the difference in overall severity between the two groups of rater expertise (New and Old) was significant and to determine whether they ranked test-takers in the same way, a Mann-Whitney U test was run. Table 6 displays the Mann-Whitney U test results between New and Old raters at the post-training phase.

Table 6.
Mann-Whitney U Test for New and Old Raters' Severity Index (Post-training)

Rater group	N	Σ Rank	Mean Rank	Std. D.	Separation	Reliability
New	10	124	12.4	0.44	1.24	0.86
Old	10	86	8.60	0.56	2.18	0.87
Z: -1.67, *p*>0.05						

As the table shows, the difference between New and Old raters at the post-training phase (Z= -1.67) is not significant (*p*>0.05), showing that the two groups of New and Old raters are not significantly different in terms of severity differences. The differences in severity between the two groups have decreased considerably which again brings evidence on the effectiveness of the training program for both groups of expertise. The range of severity estimate has reduced considerably, yet more for New raters than Old raters, by examining the standard deviation of the severity estimates which is 0.44 for New raters and 0.56 for Old ones. This reduction in severity is also reflected in the separation indices for the two groups, i.e., 1.24 for New raters and 2.18 for Old raters. The reliability index (column seven) demonstrates the reliability measure of the separation index. This measure is relatively high for both rater groups.

Through comparing the results of separation indices for both groups of raters at the post-training phase compared to that of the pre-training, a higher reduction severity was observed for New raters, showing the more constructive effect of training for New raters than Old ones. Similar to the pre-training phase, a rater-test-taker bias interaction analysis for various ranges of bias logit was performed. Table 7 displays the bias between raters (at each expertise level) and test-takers for various logit range values at the post-training phase. Over half of the interactions (127) occurred around the mean, i.e., between -0.99 to 0.99 logit values. This shows that the greatest numbers of rater-test-taker interactions were clustered in that range. Unlike the pre-training phase, New raters had a less significant bias towards test-takers than Old ones (60 to 123) which showed that New raters were a lot less biased towards test-takers than Old ones. The mean number of significant bias interactions for New raters was found 6.0 and for Old ones 12.3. This can be due to the fact that New raters benefitted more from the feedback and the training program and thus could use better strategies to judge the test-takers in accordance with their true oral ability. Rater training and the provision of feedback could affect New raters much more than Old raters in reducing their bias interaction in test-takers oral ability evaluation.

Similar to the pre-training phase, the table shows that raters have a tendency to show more bias towards higher-ability test-takers than the lower-ability ones. There were 97 bias interactions above 0.00 and 86 bias interactions below 0.00. Bias interactions for higher-ability test-takers were more likely to be severer than lenient (54 severe and 43 lenient). However, bias interactions for lower-ability test-takers were more likely to be lenient than severe (46 lenient and 40 severe). The same pattern was applicable even at the extreme ends of the scale (logit values from -3.0 to -3.99 and from 3.0 to 3.99) as well. The highest-ability test-takers attracted two out of five severe interactions, whereas the lowest-ability test-takers attracted three out of four lenient interactions. The reason again might be because the raters' expectations of test-takers rise as the ability of test-takers increases, thus making their judgments severer, too. For lower-ability test-takers, perhaps raters would benefit test-takers to compensate for their lack of proficiency. However, the result suggests that the training program, in both extreme ends of the scoring continuum, was effective in reducing the extreme ratings by the raters even at those points of the scoring scale.

Table 7.
Bias Interaction Frequency between Raters at Each Expertise Level and Test-takers for Various Bias Logit Range (Post-Training)

Logit band ➔ Rater 🡇	-3.0 to -3.99		-2.0 to -2.99		-1.0 to -1.99		0.0 to -0.99		0.0 to 0.99		1.0 to 1.99		2.0 to 2.99		3.0 to 3.99		Total (Scale in whole)
Severity/Leniency	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	
New raters	0	1	0	1	2	3	10	13	8	13	1	4	0	2	0	2	60
Old raters	1	2	3	3	6	5	18	18	29	18	8	3	6	0	2	1	123

Mean rater-test-taker bias interaction for:
New raters: 6.0
Old raters: 12.3

Discussion

The research question was related to differences between the raters of the two groups of expertise in terms of severity, bias, and consistency measures. The findings showed that the raters achieved higher measures of consistency and reduced measures of severity after training. Such a finding confirms the constructive impact of the training program in reducing raters' biases. On the other hand, the findings demonstrated that the training program was more effective for New raters than Old ones since New raters achieved higher measures of consistency and reduced measures of bias than Old ones. This finding is in line with some previous studies (Ahmadi & Sadeghi, 2016; Attali, 2016; Bijani, 2010) which found that training programs reduced raters in raters' scoring. One of the predictions made was that New raters would have a wider range of severity estimates before the administration of the training program than Old ones; however, this difference would be reduced after training. These predictions were rather supported by the above-mentioned data; at the pre-training phase, the separation indices were found 4.63 and 2.69 (See Table 1) for New and Old raters, respectively, whereas at the post-training phase, such measures were reduced to 1.89 and 2.47 (See Table 5) respectively. By a look at these results, we can observe a trend towards equilibrium between the two groups of raters. Besides, New raters tended to move closer to each other in biasedness than Old ones. However, it should be noted that due to the small sample size, the difference in data variation with respect to severity estimates must be interpreted with caution.

In summary, New and Old raters, by virtue of having systematically built up their proficiency, seem to have different perceptions from each another and, of course, in their judgment of test-takers' performances. New raters tended to display significantly more leniency and extreme negative bias (leniency) as well as inconsistency than Old raters before training. This indicates that Old raters used the scoring rubric more strictly than New ones. This finding is rather consistent with some previous research (e.g., In'nami & Koizumi, 2016), but in contrast with some other studies (e.g., Davis, 2016) which found that inexperienced raters were severer in their ratings than experienced ones. Old raters seem to have been less tolerant of test-takers' mistakes and that is why they were rather harsh to them at the pre-training phase. New raters appeared to have favored advanced test-takers as well and awarded them high scores before training due to the fact that they had received more credits. This is exactly what is known as halo effect (McNamara, 1996) which negatively affects raters' performances.

However, after training, New raters almost totally removed leniency in their ratings. New raters proved to be more likely to be within the limits of acceptability following training and tended to be more reluctant to award extremely low scores to weak candidates, whereas for Old raters, although they reduced harshness to a considerable extent and moved towards higher consistency, the training does not seem to be as effective. Also, New raters' bias and inconsistency were reduced a lot more than Old raters after training which demonstrated that they benefited more from the feedback and training than Old ones. Such finding provides further evidence on the outcome of the previous research (Barkaoui, 2011; Bijani, 2010; Galloway, cited in Bonk & Ockey, 2003; Kim, 2011) which found that the consistency of New raters improved much more after training compared to Old raters. Nevertheless, this finding was in contrast to that of Lim (2011) who found a relationship between consistency in rating and frequency of rating, i.e., Old raters were identified to be more consistent due to the higher frequency in scoring test-takers' performances. It is also noteworthy to indicate that Van Moere (2012) in his study on Chinese EFL learners found contradictory outcome; no significant difference was observed between two groups of raters with regard to the degree of severity.

The results of the study demonstrated that New raters tended to be more lenient in the majority of the rating scale categories than Old ones. This finding is consistent with that of Kuiken and Vedder (2014) who found that New raters were significantly more lenient in their ratings of coherence and fluency, and by Davis (2009), who found that Old raters were significantly harsher in their ratings than New raters in the holistic scores of speaking ability, particularly when scoring fluency. Meanwhile, it must be noted that the obtained results are fairly contradictory to that of Ahmadi and Sadeghi (2016) who found that New raters tended to be severer than Old ones with respect to the scoring of test-takers' pronunciations who, mostly, overfitted the model. On the other hand, New raters' were more biased in scoring Grammar throughout the entire study which indicates their higher concentration on this category of the rating scale, a result fairly in line with that of Kyle et al. (2016) and against that of Barkaoui (2011) who found contradictory results regarding the degree of attention paid by Old raters.

The results showed that ,while Old raters tended to concentrate more on the form of language speech production, New raters focused on its content. The results revealed that Old and New raters employed different rating approaches depending on scale descriptors and features of language, which was in line with previous findings in the field. For instances, Kim (2015) found that Old raters tended to emphasize more on the communicative aspect of language when rating students' oral performances, whereas for New raters, the focus was more on pronunciation which was reflected in the raters' produced verbal protocols, too. The findings of data analysis demonstrated that New raters can rate as reliably as or even much better than Old raters. Also, regarding the assessment criteria, the use of a rating scale along with its descriptors will result in more accurate and consistent scoring by even New raters. Such finding is parallel with the one found by Davis (2016) who argued that training program and feedback can result in higher measures of consistency among raters. Although the protocol analysis displayed that Old raters provided more comments and detailed ones, the final outcomes showed no significant qualitative difference between them. Therefore, the results offered no evidence based on which New raters should be excluded from rating solely because of their lack of adequate experience.

This finding, which is in line with that of some studies (Bijani, 2010; Nakatsuhara, 2011), suggests that the effects of feedback and training may be more effective for New raters than Old ones. This can be due to the fact that Old raters, because of their idiosyncratic characteristics such as arrogance or overconfidence, may be less likely to receive training and feedback from authorities. Old raters did not seem to welcome further education and that is why their rating was developed much less than New raters. New raters, based on the use of the scoring bands of the rating scale, seemed not to have the tendency to fail the test-takers nor to award advanced level proficiency rating so as to follow the assumptions of fairness, while Old raters appeared not to care about the issue and still were likely to follow their own scoring style, but with higher moderation after training. This finding is closely in line with that of some studies (Huang et al., 2016; Winke et al., 2012) which found that teachers tended to be more logical in their scoring compared to native speakers who were rather idealistic.

In general, the findings of the study revealed that both groups of rater expertise benefited from the administration of the training program and thus achieved higher measures of interrater reliability accordingly. The findings are parallel with those of several studies (Attali, 2016; Bijani & Fahim, 2011; Davis, 2016; Khabbazbashi, 2017) which indicated that extremely severe or lenient New raters benefited from the training program thus modified their rating behavior, making it like that of other raters.

Conclusion

This study added more proof to the usefulness of MFRM in analyzing the sources of variability in oral assessment because of raters' biases. MFRM provides more validity by removing rater variability in assessing students' performance ability. This can definitely contribute to the test fairness and accuracy of oral performance assessment. The findings showed that it is almost impossible to completely eradicate rater variability even through rater training. This shows that rater variation is a substantial element of rating. Therefore, rater training should be viewed as a procedure to bring raters as close as possible in terms of rating language performances. Besides, rater training procedure, unlike making raters consistent with each other (interrater reliability), had better make them more self-consistent within themselves (intrarater reliability).

Similar to previous research, the findings demonstrated that experienced raters, due to their idiosyncratic characteristics, did not benefit as much as inexperienced ones. Also, some amount of severity was still left after training, which may have an impact on future interpretations and decisions. This is something that could be better achieved through more training and individual feedback but not thoroughly removed. The outcomes of fit statistics analysis of the raters demonstrated that raters tended to increase their internal consistency in ratings through receiving training, feedback, and gaining experience. This shows that the facet of rater does not always represent a validity-threatening aspect of assessment, while some other facets have contributing effects. The training program resulted in the reduction of raters' biases to the rating scale categories; however, this reduction was more significant for inexperienced raters than the experienced ones, confirming the more constructive impact of training on inexperienced raters. Finally, raters' of both groups of expertise achieved higher measures of interrater reliability after training. However, it was the inexperienced raters who achieved higher measures of consistency than experienced ones.

Inexperienced and experienced raters' rating quality improved as a result of training. However, inexperienced raters were the ones who achieved much more improvement than the other group and benefited more from training. Similarly, the study showed that New raters can rate as reliably as or even much better than experienced raters. Therefore, the results offered no evidence based on which inexperienced raters should be excluded from rating solely because of their lack of adequate experience and advocating the recruitment of only experienced raters for the sake of higher reliability. Inexperienced raters, being more economical than the experienced ones, cost less for decision-makers to perform the rating task. They also showed to be more reliable after training or even without training, if standards are met. Although it is a general belief for decision-makers to select experienced raters for achieving higher reliability, the finding showed the reverse. Therefore, instead of charging a bulky budget on experienced raters, decision-makers had better use the budget for establishing better training programs. Consequently, there is no reason based on which to exclude inexperienced raters from rating. Since this study used Iranian raters, further research could be conducted using raters of other nationalities and other contexts. Besides, future studies can apply various tasks, other than the ones used in this study, for test-takers' oral performance assessment.

References

Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. *Language Assessment Quarterly, 13*(4), 341-358.

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

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.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49-58.

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.

Caban, H. L. (2003). Rater group bias in speaking assessment of four L1 Japanese ESL students. *Second Language Studies, 21*(1), 1-44.

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

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367-396.

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, Germany: Peter Lang Edition.

Educational Testing Service (2001). *ETS oral proficiency testing manual*. Princeton, NJ: Author.

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

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

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.

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

Kim, H. J. (2011). *Investigating raters’ development of rating ability on a second language speaking assessment* (Unpublished doctoral dissertation). University of Columbia, New York.

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

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31.

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.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543-560.

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

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

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

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

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325-344.

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.