THE JOURNAL OF ASIA TEFL
Vol. 14, No. 4, Winter 2017, 687-702
http://dx.doi.org/10.18823/asiatefl.2017.14.4.7.687

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Investigating the Effect of Training on Raters' Bias toward Test Takers in Oral Proficiency Assessment: A FACETS Analysis

Houman Bijani
Islamic Azad University, Iran

Mona Khabiri
Islamic Azad University, Iran

Typically, variability among raters in scoring and their bias is mediated through rater training. However, questions still remain about whether training can affect raters' severity or leniency. Furthermore, few studies have looked at the differences between trained and untrained raters in oral assessment. Oral test scores of 200 test takers rated by 20 raters and were analyzed before and after a training program using the multifaceted Rasch measurement (MFRM). The results demonstrated the constructive impact of training programs in reducing raters' biases and increasing their consistency measures. This study indicated that inexperienced raters benefited more from a training program than experienced raters and thus achieved higher measures of consistency afterward. It also demonstrated a higher biased interaction for test takers on the extreme ends of the oral ability continuum. The findings demonstrated that it is almost impossible to completely eradicate rater variability even through rater training. Therefore, rater training should be viewed as a procedure to establish within-rater consistency rather than between-rater consistency. Since this study showed that inexperienced raters can rate even more reliably than experienced ones after training, there is no evidence whereby decision makers can exclude inexperienced raters solely because of their lack of adequate experience. Consequently, decision makers need to use their budgets for establishing rater training programs for inexperienced raters instead.

Keywords: bias, feedback, interrater reliability, intra-rater reliability, multifaceted Rasch measurement (MFRM), rater severity/leniency, rater training

Introduction

The chief concern of oral performance assessment is to evaluate test takers' substantive speaking ability 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) emphasize that test takers are not separated figures that are only responsible for their performance, while the interaction among factors such as interlocutors, raters, and test versions also influence test takers' performance. Bachman and Palmer (1996) also believe that there are factors that influence language test scores and stressed the understanding of these factors to diminish their influences on test performance.

Although it is impossible to entirely eradicate the influences of these variables on test takers' performance, it is essential to identify and control them to better evaluate test takers' speaking ability and

more suitably interpret and use the test outcomes (Gan, 2010). A number of possible sources of rater disagreement have been studied and explored in the literature on speaking assessment in both first and second language contexts (e.g., Fulcher, 1994; Huang, Huang, & Hong, 2016; McNamara, 1996; Trace, Janssen, & Meier, 2017). One of the most prominent concerns of test score variability sources, no matter how carefully the test is constructed, is the issue of raters. Traditional theories of performance assessment have regarded rater characteristics in terms of the difference between an idealized rater and an actual rater. Here, the problem is with the actual rater which is called error. This difference or error can be conceptualized in terms of severity or leniency (Kim, 2015; Lunz & Stahl, 1990). McNamara (1996) asserts that differences between raters are varied. One rater may be more lenient than another or a rater may be more lenient or severer to particular test takers, groups, specific tasks, various testing contexts, or situations. The interaction between the rater and these various aspects of assessment is referred to as bias. According to Lumley and McNamara (1995) there are two types of bias: Type I (leniency bias) in which test takers undeservedly are awarded higher score than what they deserve, while in type II (severity bias) they receive lower scores than what they deserve. The issue at the heart of increasing both measures of reliability and validity in performance assessment (e.g., speaking assessment) and reducing both types of bias turns out to be rater training (Chalhoub-Deville, 1995b)

Literature Review

Raters' Background in Oral Performance Assessment

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 backgrounds and the criteria they apply (Barrett, 2001). Several studies have found differences between inexperienced and experienced raters in their scorings and their use of rating strategies (Khabbazbashi, 2017; Nakatsuhara, 2011; Winke & Gass, 2013). Khabbazbashi (2017) compared the ratings of trained and untrained raters from two various backgrounds: experienced English teachers and non-teachers. Half the raters from each background group received training and the other half did not. 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. Commonly, attempts are made to reduce variability among raters' behaviors related to their differences in scoring and their bias through rater training (e.g., Attali, 2016; Bijani, 2010; Davis, 2016). This is done in rater training sessions by introducing the raters to the assessment criteria and asking them to rate already scored performances representing various language proficiency levels. Raters are then provided with feedback on the extent to which they are in line with other raters (Zhang & Elder, 2011). In speaking, rater training is used to modify raters' expectations of tasks and test takers' characteristics (Elder, Barkhuizen, Knoch, & Randow, 2007), and to clarify various elements of the rating scale in order to reduce levels of rater variability (Xi & Mollaun, 2011).

A number of research studies have shown that rater training minimizes rater effects (e.g., Attali, 2016; Bijani, 2010; Davis, 2016; Khabbazbashi, 2017; Winke & Gass, 2013). Winke and Gass (2013) found that providing feedback on raters' scoring behaviors could assist them in becoming more consistent in subsequent ratings and rater training. Rater training can help raters better understand the categories and criteria of the rating scales which might influence their rating behavior (Kuiken & Vedder, 2014), and Bijani (2010) found training effective in improving rater reliability. Accordingly, in the absence of rater training programs, raters with various levels of experience may assign different scores to the language being tested (Attali, 2016; Bijani, 2010; Davis, 2016), while extended training programs will aid them develop a common reference framework. There is evidence that rater training can be effective, in that although it cannot totally eliminate extreme differences, it can reduce variability in rater severity and increases the self-consistency of raters by reducing individual biases of raters. That is, although rater

training cannot totally eliminate rater variability, it can make raters more self-consistent (e.g., Bijani, 2010; Bijani & Fahim, 2011; Davis, 2016; Elder, Barkhuizen, Knoch, & Randow, 2007; Khabbazbashi, 2017; Nakatsuhara, 2011).

Rater Behavior in Oral Performance Assessment

One more source of rater variation is how the behavior of the rater or the interviewer which can directly influence the outcome of performance assessment. Some previous research on rater behavior has demonstrated a considerable amount of rater variability, which is mostly related to raters' characteristics and not the test takers' performance (e.g., Carey, Mannell, & Dunn, 2011; Knoch, 2011). Accordingly, several research studies have been carried out to investigate the accuracy of oral performance assessment tests through research on both rater reliability, which is the degree of agreement among independent raters in assessing the test takers (Khabbazbashi, 2017; Winke & Gass, 2013), and the validity of performance-based tests through construct validity and concurrent validity (Kyle, Crossley & McNamara, 2016). Furthermore, rater training has demonstrated low impact in reducing this variability. That is, previous research studies have shown variability among raters even after extensive rating programs. There are various manifestations of rater variability: (1) the degree to which they agree with the rating scale; (2) how they interpret the criteria in the rating scale; (3) the extent of raters' severity or leniency in rating; and (4) the degree to which their scorings are consistent across test takers, scoring criteria, and performance tasks (Kim, 2011).

A number of studies have investigated the impact of rater variability in rating test takers' oral performance. Davis (2016) had 51 raters grade interviewer behavior using a questionnaire. The results of the FACETS analysis, computer software used for the multifaceted Rasch measurement, revealed that the raters identified interviewers as "good" if they could establish rapport with their test takers. Also, the best ones were those that used significantly more markers of politeness and were actively involved in the interview. Winke, Gass, and Myford (2012) analyzed the oral performance of 32 IELTS test takers as judged by six interviewers. Those performances which were rated differently were reanalyzed and the outcomes showed that "easy" interviewers used more common topics, were more friendly, and used more words of encouragement, whereas the "difficult" interviewers challenged the candidates more and acted more like examiners.

Attali (2016) studied four inexperienced raters scoring compositions both before and after the rater training program. The outcome demonstrated that clarification of scoring criteria, modification of their awareness, and awareness of the need for inter-rater agreement consequently brought inexperienced raters more in line with the experienced ones. Kim (2011) used MFRM in a research study on group oral assessment for Japanese EFL learners, and used a group of expert and non-expert raters to assess the test takers' spoken language. The results demonstrated high variability among raters, and it was also found that raters tended to become stricter with experience. Van Moere (2012) compared the performance of experienced and inexperienced English raters scoring Chinese students studying English in the US. The findings showed no significant difference between the two groups of raters in their degree of severity. Barkaoui (2011) classified raters into three groups of experience, and with the help of verbal protocols, they found qualitative differences in how 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, there still existed some qualitative differences in their rating approaches, which were attributed to their experience.

The use of MFRM in bias analysis yields several implications in performance assessment. First, MFRM helps researchers study rater facet (variable) with regard to their facet of interest by keeping the other facets constant and neutral (Wright & Linacre, 1994). Second, it can help researchers administer rater training programs. Research has shown that rater consistency and rating validity can be increased through training (Kyle, Crossley, & McNamara, 2016). Third, MFRM can help reduce self-inconsistency

and increase intra-rater reliability, which increases the fairness of a test, specifically in placement and summative evaluation tests (Gan, 2010). In another study, Lumley and McNamara (1995) investigated three sets of graded spoken English tests over a period of 20 months. The findings of the interactional effects of time and rater facets represented a significant change in rater's severity.

However, little is known about what actually happens during rater training and how it affects the raters' change in behavior in scoring. It is still unclear whether training can affect raters' severity/leniency, since, according to Barrett (2001), rater severity is a stable and unchangeable rater characteristic which varies from rater to rater. Regarding bias, most studies conducted so far (e.g., Bijani & Fahim, 2011; Kim, 2011; Kondo-Brown, 2002) have not addressed the interaction of raters' severity/leniency with test takers' ability facets. While a few studies have looked at the differences between trained and untrained raters in speaking assessment (Bijani, 2010; Elder, Barkhuizen, Knoch, & Randow, 2007; Gan, 2010; Kim, 2011), few if any, studies have used a pre- and post-training design. Although a few studies have investigated the influence of training in second language speaking assessment (e.g., Barrette, 2001; Davis, 2016; Saito, 2008), they have not provided enough conclusive evidence about the impact of the training programs on raters' severity/leniency, or bias and consistency measures. Also, raters do not agree with each other or even within themselves on scoring any oral performance. More importantly, even if raters assigned exactly the same score to a test taker, there would still be vague points about the interpretability of those scores.

Considerable evidence of poor rater consistency has been reported in some research (e.g., Lunz & Stahl, 1990; Trace, Janssen, & Meier, 2017), and even if adequate consistency might have been reported in most research, it is mostly on the basis of correlations alone. That is, even a perfect correlation might ignore systematic variations among raters. Therefore, the purpose of this study is to investigate the amount of raters' biases in relation to test takers' oral ability before and after a training program, and additionally, to what extent the administration of the training program would affect raters' biases in their subsequent ratings over both short and long terms. The long-term analysis will determine to what extent the training program would help raters minimize their biases more than before training. Consequently, the following research question can be formulated:

RQ: Is there any significant difference in rater-test taker biased interaction after the training program than before training?

Methodology

Participants

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

20 Iranian EFL teachers, including 10 males and 10 females, ranging in age from 24 to 58 were selected randomly among those who volunteered to participate in this study as raters. These raters were undergraduates and graduates in English language related fields of study, teaching in different universities and language institutes. It should also be noted that all the raters had high levels of English language proficiency, although none was a native speaker of English. In order to fulfill the requirements of this study, the raters had to be classified into two groups of experienced and inexperienced raters to investigate the similarities and differences among them and the likely advantages of one group over the other. Each rater was assigned a number from 1 to 10 to protect their anonymity and confidentiality of their data. In order to search for rater participants for the present study, a background questionnaire, adapted from Lumley and McNamara (1995), was given to the raters, eliciting the following information: (1) demographic information, (2) rating experience, (3) teaching experience, (4) rater training and (5)

relevant courses passed. Based on these methods of rater classification, these 20 raters were divided into two levels of experience, each containing 10 raters, according to their experience as 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 who had passed less than the four core courses related to the ELT major (pedagogical English grammar, phonetics and phonology, second language acquisition, and second language assessment). Hereafter we designate these raters as new raters, who were 10 in number.
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 who had passed all the four core courses plus at least two elective courses related to the ELT major. Hereafter we designate these raters as old raters, who were 10 in number.

Instruments

Oral tasks

The present study used the Community English Program (CEP) test to evaluate test takers' speaking ability under various language use situations. The purpose of the speaking test was to measure the extent to which second language speakers could produce meaningful, coherent, and contextually appropriate responses to the five tasks. Task 1 (*Description Task*) was an independent-skill task which reflected test takers' personal experience or background knowledge to respond in a manner for which no input was provided (Bachman & Palmer, 1996). For tasks 2 (*Narration Task*) the test takers were required to respond to pictorial prompts based on a sequence of pictures. On the other hand, tasks 3 (*Summarizing Task*) and 4 (*Role-play Task*) reflected test takers' use of their listening skills to respond orally. That is, a prompt was provided for the test takers through listening _ listening to short or long prompts. In task 5 (*Exposition Task*), the test takers were required to describe, explain, analyze, and interpret information in the form of tables, graphs or diagrams. The speaking tasks were administered through face-to-face oral interactions between the test takers and the raters of around 15 to 20 minutes. Each interview interaction was audio-recorded and then assessed by raters. As it was reiterated before in the literature review, since face-to-face interaction was commonly considered to be the main context for speaking in the real world, the direct method of delivery was quite attractive because of the high amount of face validity it offered to test developers, test takers and test users.

Scoring rubric

As one of the requirements of this study, in order to evaluate the influence of using a scoring rubric on the validity and reliability of assessing test takers' oral performance, this study aimed to make use of an analytic rating scale. The purpose of using an analytic rating scale was to assess test takers' oral performance in order to determine to what extent it evaluates the oral proficiency of test takers in a more valid and reliable way. Each test taker's task performance was assessed by both new and old raters using the ETS (2001) analytic rating scale. In the ETS (2001) scoring rubric, individual tasks are assessed using appropriate criteria including fluency, grammar, vocabulary, intelligibility, cohesion and comprehension. Each of these criteria is accompanied by a set of 7 descriptors. All scoring is done on a Likert scale from 1 to 7.

Procedure

Pre-training phase

Prior to collecting any data from the test takers, the raters' background questionnaire was given to the raters to fill out before starting the test tasks. The raters were not informed about which group of expertise they belonged to, because this might cause some inexperienced raters to feel intimidated by the experienced ones. The 200 test takers were divided randomly into two groups such that each group took part in each phase of the study (pre- and post-training). Half the data, i.e., 100 recordings, were collected from the test takers in the pre-training phase, and the rest in the post-training data collection stage. These 100 test takers who took part in the pre-training phase took all the five oral tasks. During the data collection exam session, the raters were given a guide to oral data collection, which clarified what they were supposed to do in eliciting oral data from the test takers in a step by step approach for the entire tasks. All the raters were given one week to submit their scorings, based on the six band analytic rating scale, to the researcher.

Rater training

After the pre-training scoring stage, 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 raters were provided with information specifically related to scoring procedures, since the aim of the training program was to familiarize all raters of various levels of expertise. Although the training session was the main component of the rater training program, it was, however, accompanied by rating norming practice, group discussion, and score negotiation. These procedures were continued until they reached consensus and all raters were confident in determining test takers' scores across the descriptors of the scoring rubrics. Therefore, the phrase "training process" refers both to formal training received in the norming session, and the informal training through socialization. Previously recorded responses were played and the raters scored them using the scoring rubric criteria under the guidance and assistance of the trainer. The training program consisted of rater norming and feedback on previous rating behavior (pre-training stage), and was conducted in two separate group norming sessions, each lasting for about six hours, with an interval of one week.

The feedback provided to the raters included the raters' previous rating patterns determined by a statistical analysis of the analytic rating scores that each rater gave (i.e., severity, internal consistency, and biases with certain test taker groups, task, and rating scale). For feedback on raters' biases, raters with z-scores beyond ±2 were considered to have a significant bias and were reminded individually to be aware of these issues accordingly. For feedback on raters' consistency, raters with infit mean squares beyond the acceptable range of 0.6 to 1.4, as suggested by Wright and Linacre (1994), were considered as misfitting, that is raters with an infit mean square value below 0.6 were too consistent (overfitting the model) and those with an infit mean square value above 1.4 were inconsistent (underfitting the model). Therefore, this was pointed out to these raters individually if they were identified as misfitting.

Post-training phase

Immediately after the training program for the raters, the oral tasks were conducted one by one for the test takers. As mentioned before in the pre-training data collection procedure, the second half of the test takers (including 100 students) were used, from whom data were elicited. Similar to the pre-training phase of the study, the 100 test takers who took part in the post-training phase took all five oral tasks. Once again, all the raters were given one week to submit their scorings, based on the six band analytic rating scale, to the researcher.

Data Analysis

In order to investigate the research questions, the researcher employed a pre/post-method research design in which a series of quantitative approaches were used to investigate the raters' development over time with regard to rating second language oral performance assessment (Cohen, Manion, & Morrison, 2007). Quantitative data (i.e., raters' scores based on an analytic scoring rubric) were collected and analyzed with MFRM before and after the rater training program. The 20 raters scored all five oral task performances by the 200 students before and after the training program. The scoring patterns of the two groups of raters 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 at each rating point, and (2) within each rater group to investigate the development of the raters' ability.

Results

A bias analysis was conducted to investigate the rater-test taker interactions. The data analysis, out of 2000 interactions (interactions between 20 raters and 100 test takers in the pre-training phase) revealed 644 significant biased interactions in the pre-training phase. Among these, 327 interactions were toward severity (positive logit values) and the remaining 317 interactions were toward leniency (negative logit values). It can thus be seen from Table 1 that there was more bias against the test takers than in favor. That is, biases were likely to be type II, where test takers undeservedly received lower scores, than type I, where they undeservedly received higher scores. The minimum number of rater-test taker biased interactions was for rater old3 with 19 biased interactions, and the maximum was for rater new6 with 48 biased interactions. Table 1 shows the rater-test taker interactions in the pre-training phase. Columns one (Rater) and two (Test taker) show a few of the raters and test takers the interaction among whom was significantly biased.

The third column (Obs-Exp scores) shows a test taker's observed score minus his/her expected score. As an example, the observed and expected score for test taker 11 were 18 and 28, respectively, out of 42 points; the difference would then be (18-28 = -10). Because the scale used in this study had six categories (fluency, grammar, vocabulary, intelligibility, cohesion and comprehension), the average difference between the observed score and the expected score would be -10/6 = -1.66 for these categories. This indicates that the test taker's score from rater old8 was much lower than the expected score; thus, rater old8 scored test taker 11 more severely than expected.

The fourth column (Bias logit) demonstrates the bias value, representing raters' severity/leniency in the performance assessment of test takers. Positive values represent severity, while negative ones represent leniency. The fifth column (SE) displays the standard error of bias estimation. The small amount of SE provides evidence for a high precision of measurement. The sixth column (z-scores, or standardized fit statistics) shows the test version rater bias estimate at this phase. Bias is the difference between expected and observed ratings of the obtained data, which is then divided by its standard error to derive the z-score (Lunz & Stahl, 1990). The most preferable z value is 0, which indicates that the data match the expected model, and thus, no rater bias. According to McNamara (1996), z values between ±2 are considered an acceptable range of bias, and thus, any values beyond this range are considered to be either too positively biased or too negatively biased.

The seventh and eighth columns (infit and outfit mean square) provide the fit statistics, which show to what extent the data fit the Rasch model, or the difference between the observed and expected scores. 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). That is, fit statistics 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. Fit statistics are categorized into two subcomponents entitled infit and outfit statistics, and most researchers employ them because they are

considered less sensitive to sample size.

Infit is the weighted mean square statistic, which is weighted toward expected responses and thus sensitive to unexpected responses near the point where the decision is made. I.e., 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 have the expected value of 1 and a range of zero to infinity; however, there is no straightforward rule, absolute or universally definite range for interpreting fit statistics values or for setting upper and lower limits; therefore, the acceptability of fit is done on a judgmental basis, not solely on a statistical one. The acceptable range of fit statistics, although varying among statisticians, according to Wright and Linacre (1994), is within 0.6 to 1.4 logit values. Therefore, in order to investigate the fit statistics value, raters below this range are overfit or overly consistent, and those above this range are underfit (misfit) or overly inconsistent.

FACETS also provides for test takers' separation statistics. The importance of separation statistics derives from their ability to provide scores that can separate test takers into specific categories (Eckes, 2015). Here, the test takers' separation index was 7.50, showing that the test takers could be classified into seven and a half statistically distinct levels of proficiency. Statistically significant distinct levels are defined as those that are 3 standard errors apart (Linacre, 2002). The fixed chi-square value for the 644 rater-test taker interactions was measured as well. The chi-square value indicates whether there was a significant difference in the interaction between the raters and the test takers with respect to raters' severity toward test takers' performance ability before training ($X^2_{(1999,\ N=2000)} = 28678.20$, $p<0.00$). The outcome suggested that the raters were not at the same level of severity. Additionally, the reliability of this separation index was measured 0.89, showing that the test takers were very reliably separated in their levels of proficiency.

TABLE 1
Rater-Test Taker Bias Analysis Report (Pre-training)

Rater	Test taker	Obs-Exp Average logit	Bias logit	SE	Z-score	Fit statistics	
						Infit Mn Sq	Outfit Mn Sq
Old3	8	-0.74	1.19	0.05	2.12	0.8	0.8
Old3	6	-0.82	1.31	0.05	2.21	0.9	1.0
Old8	11	-2.03	3.24	0.13	4.87	1.9	1.9
Old8	18	-1.66	2.57	0.10	3.86	1.5	1.5
Old8	25	-1.12	1.79	0.07	2.68	1.1	1.2
New10	35	1.88	-3.01	0.12	-4.52	0.6	0.6
New10	71	1.09	-1.75	0.07	-2.62	1.0	1.1
New10	66	0.88	-1.40	0.06	-2.10	0.8	0.9
New10	19	1.18	-1.88	0.08	-2.82	1.1	1.1
Old7	20	1.02	-1.63	0.07	-2.45	1.0	1.0
Old7	92	0.96	-1.54	0.06	-2.30	0.9	0.9
Old7	78	1.00	-1.59	0.06	-2.39	1.0	1.1
New1	81	1.03	-1.65	0.07	-2.48	1.0	1.0
New1	85	1.13	-1.80	0.07	-2.71	1.1	1.1
New5	84	1.08	-1.73	0.07	-2.59	1.0	1.1
New5	34	1.27	-2.04	0.08	-3.05	1.2	1.2
New5	61	1.18	-1.88	0.08	-2.82	1.1	1.1
New6	69	1.40	-2.25	0.09	-3.37	1.3	1.4
New6	48	1.69	-2.71	0.11	-4.06	1.6	1.6
New6	52	1.92	-3.07	0.12	-4.61	1.8	1.8
New6	4	2.21	-3.53	0.14	-5.30	0.3	0.3
New6	33	0.95	-1.52	0.06	-2.28	0.9	0.9
Mean		0.70	-1.13	0.08	-1.66	1.08	1.10
SD		1.17	1.87	0.02	2.84	0.41	0.38

Test takers' separation index: 7.50
Fixed (all-same) chi-square: 28678.20, *df*= 1999, $p<0.00$
Reliability: 0.89

In order to obtain a better demonstration of the systematic pattern of rater-test taker biased interactions, a rater-test taker biased interaction analysis for various ranges of the bias logit was performed. Table 2 displays the rater-test taker biased interactions for various logit range values at the pre-training phase. The mean number of significant biased interactions for each rater was 32.2, which indicates that each rater had 32.2 severe and/or lenient biased interactions with the test takers when rating their oral performances on the 5 tasks. Over half of the interactions (361) occurred around the mean, i.e., between -0.99 to 0.99 logit values. This can be due to the fact that the greatest numbers of rater-test taker interactions were clustered in that range. The table displays rater-test taker interaction frequencies based on logit ranges, and shows that raters have a tendency to show more bias toward higher ability test takers than lower ability ones (327 severities and 317 leniencies). There were 332 biased 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), while biased interactions for lower ability test takers were more likely to be lenient than severe (171 lenient and 141 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 12 out of 18 severe interactions, while the lowest ability test takers attracted 9 out of 13 lenient interactions. This indicates that the raters demonstrate more severity in rating when rating highly competent test takers; however, they were fairly lenient in their ratings toward extremely weak test takers. This finding is parallel, albeit in a writing assessment test, to one found by Schaefer (2008), who in an analysis of ratings by 40 native English speakers of 40 essays by Japanese students found some raters scored higher ability test takers more severely and lower ability ones more leniently than expected. The reason of this interaction tendency is not quite clear; however, it might be due to the fact that raters' expectations of test takers rise as test takers' abilities increase, thus making their judgments severer. For lower ability test takers, perhaps raters would favor 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 bias, but in writing tasks. Interestingly, at this phase, the new raters had more significant bias toward test takers than old ones (337 to 307), showing that old raters were less biased toward test takers than new ones. This may be because old raters used better strategies as a basis for judging test takers on their true oral ability.

TABLE 2
Rater-Test Taker Biased Interaction for Various Bias Logit Ranges (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
Severity/Leniency	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	
New1		1		1	1	4	4	1	3	5	3	1	1	2			27
New2		1		3	3	1	1	13	1	8	1	2	3	3		2	42
New3			3		1	4	7	2	9	3	4		3		1		37
New4			1	2	2	1	2	4	5	3	3	1	1	2			27
New5			1	2	2	3	4	4	4	4	2	1	2				29
New6	1	1		4	3		4	11	2	11	1	5		3		2	48
New7			2		1	4	6	2	9	2	5	1	2		2		36
New8		1		1	1	2	5	4	2	6	2	1	2	1			28
New9				2	2	2	4	2	2	5	1	1		1			22
New10		1		4	2		2	11	1	11	1	5		2		1	41
Old1			3			4	8	2	11	3	5		3		1		40
Old2			1	2	2	1	5	4	5	3	3	1	1	1	1		30
Old3		1				3	4	3	2	2	1	2	1				19
Old4	1		3		1	3	8	2	9	1	4	1	2		1		36
Old5	1			1	2	1	4	3	4	4	2	1	1	1			25
Old6			1	1	1	3	3	4	4	7	2		1		1		28
Old7				4	1	2	2	8	2	9		5		2		1	36
Old8	1		4	1		6	10	2	11	1	4		3		3		46
Old9		2	1	1	1	1	2	1	4	2	3	1	1		2		22
Old10		1	1	1	1	2	4	2	5	2	4	1	1				25
Total	4	9	21	30	27	47	89	85	95	92	51	30	28	18	12	6	**644**

Mean rater-test taker biased interaction: 32.2

A second bias analysis was also conducted to investigate the rater-test taker interactions at the post-training phase. The data analysis of 2000 interactions revealed 183 significant biased interactions after training. Among these, 94 interactions were toward severity (positive logit values) and the remaining 89 interactions were toward leniency (negative logit values). Similar to the pre-training phase, albeit with less difference, it can be seen from the Table 3 that there were more bias terms against the test takers than in their favor, i.e., type II bias. The minimum rater-test taker biased interactions were for rater new8, new9 and new10 with 2 bias interactions, and the maximum was for rater new6 with 22 biased interactions.

Once again the FACETS was used to obtain the test takers' separation index. The test-taker separation index measured 6.78, which showed that the test takers participating in this study could be classified into almost seven statistically distinct levels of proficiency. The fixed chi-square value for the 183 rater-test taker interactions was measured as well; the chi-square value indicated whether there was a significant difference in the interaction between the raters and the test takers with respect to raters' severity toward test takers' performance ability after training ($X^2_{(1999, N=2000)} = 9854.52$, $p<0.00$). The outcome suggested that the raters were not at the same level of severity. On the other hand, the reliability of this separation index was measured 0.84, showing that the test takers were rather reliably separated into various levels of proficiency according to their scoring by the raters. Table 3 shows the rater-test taker bias analysis at the post-training phase.

TABLE 3
Rater-Test Taker Bias Analysis Report (Post-training)

Rater	Test taker	Obs-Exp Average logit	Bias Logit	SE	Z-score	Fit statistics	
						Infit Mn Sq	Outfit Mn Sq
Old8	41	-1.15	0.72	0.08	1.38	1.1	1.1
Old8	22	-0.86	0.54	0.07	0.92	0.7	0.6
Old8	67	-2.27	1.42	0.12	2.49	1.6	1.6
Old4	9	-1.81	1.13	0.06	2.11	1.2	1.3
Old4	13	-1.10	0.69	0.18	1.32	1.1	1.2
Old4	28	-1.94	1.21	0.13	2.06	1.2	1.2
New3	95	-0.18	0.11	0.09	0.19	0.7	0.8
New3	50	-0.99	0.62	0.14	1.05	0.9	0.9
New8	44	-0.43	0.27	0.08	0.42	0.8	0.9
New8	82	0.13	-0.08	0.11	-0.14	0.1	0.1
New8	35	-0.26	0.16	0.06	0.27	0.6	0.6
Old5	46	-0.30	0.19	0.13	0.32	0.7	0.9
Old5	71	0.38	-0.24	0.11	-0.41	0.7	0.7
Old2	53	0.21	-0.13	0.10	-0.22	0.7	0.8
Old2	84	0.77	-0.48	0.12	-0.82	0.7	0.7
Old3	17	1.23	-0.77	0.09	-1.31	1.0	1.1
Old3	68	1.36	-0.85	0.07	-1.44	1.2	1.2
Old3	27	1.33	-0.83	0.08	-1.41	1.1	1.2
Old9	91	1.23	-0.77	0.06	-1.31	1.0	1.0
Old9	100	1.66	-1.04	0.05	-1.77	1.3	1.3
New6	77	1.97	-1.23	0.04	-2.09	1.4	1.2
New6	59	1.50	-0.94	0.12	-1.60	1.3	1.2
New6	29	1.06	-0.66	0.11	-1.12	0.9	1.0
Mean		0.67	-0.40	0.09	-0.40	0.94	0.98
SD		1.24	0.78	0.03	1.35	0.35	0.31

Test takers' separation index: 6.78
Fixed (all-same) chi-square: 9854.52, *df*= 1999, $p<0.00$
Reliability: 0.84

Once again, in order to better demonstrate a systematic pattern of rater-test taker biased interactions, a rater-test taker biased interaction analysis for various bias logit ranges was performed. Table 4 displays the rater-test taker biased interaction for various logit range values after training. The mean number of significant biased interactions for each rater was 9.15. Similar to the pre-training phase, over half of the

interactions (127) occurred around the mean, i.e., between -0.99 to 0.99 logit values. This showed that that the greatest numbers of rater-test taker interactions were clustered in that range. The table displays rater-test taker interaction frequencies based on logit ranges. Similar to the pre-training phase, raters have a tendency to show more bias toward higher ability test takers than the lower ability ones. There were 97 biased interactions above 0.00 and 86 biased interactions below 0.00. Biased interactions for higher ability test takers were more likely be more severe than lenient (54 severe and 43 lenient). However, biased interactions for lower ability test takers were more likely to be lenient than severe (46 lenient and 40 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 attracted 2 out of 5 severe interactions, while the lowest ability test takers attracted 3 out of 4 lenient interactions. As indicated previously, although the reason for such abnormality in the raters' rating behavior is not clear, the outcome suggests that the training program, in both extreme ends of the scoring continuum, was effective in reducing extreme ratings by the raters even at those points of the scoring scale.

Unlike the pre-training phase, new raters had less significant bias toward test takers than old ones (60 to 123), which showed that new raters were much less biased toward test takers than old ones. That is, the data analysis revealed that old raters had more than twice as many biases as new raters. This may be because new raters benefitted more from the feedback and the training program and thus were able to use better strategies as a basis for judging test takers according to true oral ability. Rater training and providing feedback affected new raters much better than old raters in reducing their biased interaction in their evaluation of test takers' oral ability.

TABLE 4
Rater-Test Taker Biased Interaction for Various Bias Logit Ranges (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
Severity/Leniency	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	
New1								1	1			1					3
New2							1	3		2							6
New3					1		1			1							3
New4						1	1	2	1	3		1		1			10
New5							1	1		1							3
New6		1		1		2	2	4	2	5		2		1		2	22
New7					1		2	1	2		1						7
New8							1		1								2
New9							1		1								2
New10								1		1							2
Old1					1	1	3	2	5	1	1		1				15
Old2						1	1	2		3							7
Old3		1		1		1	1	4	1	3		1					13
Old4			1	1	2		3	2	5	2	2		1		1		20
Old5								1	2								3
Old6					1		2	1	3	1	1		1				10
Old7			1		1		3		4	1	1	1	1				13
Old8	1		1		1	1	2	1	6	1	2		2		1		19
Old9		1		1		1	1	4	1	6		1				1	17
Old10							2	1	2		1						6
Total	1	3	3	4	8	8	28	31	37	31	9	7	6	2	2	3	**183**

Mean rater-test taker biased interaction: 9.15

Table 5 displays a summary of raters' behavior in their consistency, severity and bias before and after the training program.

TABLE 5
Summary Report of Rater-Test Takers Biased Interactions (Two Phases)

Study phase	Frequency of bias	Bias direction		Min. and max. bias frequency for raters		Bias between -0.99 and 0.99 logit	Bias direction				Extreme bias frequency		Expertise bias frequency		Sig. $p<0.05$* $p<0.01$**
		Toward severity	Toward leniency	Min.	Max.		Toward higher ability test takers		Toward lower ability test takers		Above 3.99 logits	Below -3.99 logits	New	Old	
							Toward severity	Toward leniency	Toward severity	Toward leniency					
Pre-training	644	327	317	19	48	361	186	146	141	171	18	13	337	307	*
Post-training	183	94	89	2	22	127	54	43	40	46	5	4	60	123	**

Discussion

The outcomes of this study demonstrated that rating may be done without training, but for reliable rating, training is essential. The primary purpose of training is to help raters articulate and justify their scoring decisions for reliable ratings. Raters differed strongly from one another before training in severity, bias and consistency; however, after training they reduced their severity and bias to a great extent, resulting in an increase in consistency in rating. Test takers' own speaking ability is the main determining factor for the scores they receive; however, the rater factor also accounts for a substantial role in the fluctuation in their scores. Such a finding is in line with that of Kim (2017), who found that the scores awarded to test takers is more dependent to their own true oral ability.

The outcome of the research question showed that new raters had less significant bias toward test takers than old raters after training, which indicates that new raters were much less biased toward test takers than old raters. After training, the data analysis revealed that old raters had more than twice as many biases as new raters. This outcome indicated that new raters benefitted more from the feedback and the training program and thus could use better strategies as a basis for judging test takers in accordance with their true oral ability. Rater training and providing feedback could affect new raters much more than old raters in reducing their biased interaction in their evaluation of test takers' oral ability. Such an outcome is in line with some previous research (e.g., Bijani, 2010; Davis, 2016; Khabbazbashi, 2017), who similarly found a more constructive effect of training on novice raters than experienced ones in their research findings. However, such findings contradict other research findings which demonstrated superior rating performance of experienced raters than inexperienced raters even after training. For example, studies by Barakaoui (2011) and Van Moere (2012) indicated more reliable rating performance by experienced raters; experienced raters showed less bias and more consistency than novice raters after training.

One very noteworthy finding of this study was that a higher biased interaction was observed for test takers on the extreme high end of language performance ability, i.e., high ability test takers. This finding is both similar to and different from that of Kondo-Brown (2002) and Kim (2011), who found that raters showed extreme bias toward both extremes of test takers' oral ability, i.e., both high ability and low ability test takers. This finding might be because some raters' expectations of test takers would rise when rating higher ability test takers than lower ability test takers. Additionally, this finding calls for clearer rating criteria and a more extensive and comprehensive training program with a more significant focus on rating test takers at the extreme ends. On the other hand, this finding demonstrates a shortcoming of the FACETS program in correctly estimating the ability of test takers on both the extreme ends of the ability continuum. Accordingly, further research exploring raters' rating performance on test takers of extremely high and low ability is required.

A possible reason for contradictory findings between this study and the previous studies (e.g., Kondo-Brown, 2002; Nakatsuhara, 2011; Winke & Gass, 2013) could be the use of different raters in different assessment settings and different test items and tasks. For example, Kondo-Brown (2002) used only trained raters in assessing test takers' writing ability, and Winke and Gass (2013) used raters from Spanish, Chinese and Korean backgrounds. Raters' various levels of expertise could also be a determining factor in this respect as well. Nonetheless the findings of this study regarding raters' biases toward test takers are quite useful, as evaluation in Iran is typically norm-referenced, i.e., rating is done by comparison with other test takers (Farhady & Hedayati, 2009). Thus, these findings could provide testing centers and decision makers with better strategies to improve fairness in scoring. However, since this study was done on a particular group of raters and test takers and in a specific context, further research is required to provide more evidence to confirm the aforementioned findings.

The substantial rater severity/leniency differences among raters, as was also found in some previous research (e.g., Attali, 2016; Bijani & Fahim, 2011; Xi & Mollaun, 2011), have important consequences for decision makers, in that in rater training more attention and importance should be dedicated to within-rater consistency (intra-rater agreement) than to between-rater consistency (interrater agreement).

Although training programs result in higher measures of interrater agreement among raters (as found in this study), they cannot completely eradicate differences among raters. The findings of this study yield noteworthy implications which confirm the warning from Kuiken and Vedder (2014) that prohibits certifying raters on the basis of a single calibration of consistent and unbiased rating.

Conclusion and Implications

The findings of this study based on statistical MFRM outcomes demonstrated the usefulness of this analytical approach in detecting rater effects and ascertaining the consistency and variability in rater behavior in order to evaluate rating quality. MFRM can provide raters with rapid feedback on their variability and thus apply adjustments on raters' behaviors based on that feedback. Such variability, albeit very small, might have a large impact on test takers' futures. This study toads further proof of the usefulness of MFRM in analyzing the sources of variability oral assessment due to rater bias. MFRM provides more validity by removing rater variability in assessing students' performance ability. This can definitely contribute to test fairness and accuracy in oral performance assessment.

The findings demonstrate 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 type of procedure to bring raters as close as possible in rating language performance. Furthermore, rater training procedures, in contrast to making raters consistent with each other (inter-rater reliability), make them more self-consistent within themselves (intra-rater reliability). This provides enough justification for using MFRM for data analysis specifically when there are no more than two raters for rating test takers, since MFRM can account for the severity/leniency of every rater. This research provides educators with a deeper understanding of various second language oral test traits which will result in improvement of test takers' L2 oral ability. Furthermore, individual rater's scoring patterns obtained from the MFRM can be used to provide future raters in training programs with valuable feedback to improve their ratings.

The outcomes suggest that decision makers need not be concerned about raters' experience. Although decision makers commonly use experienced raters for 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. Through rater training programs, rater effects and variability can be controlled. Thus, decision makers need to establish rater training programs to increase rater consistency and reduce bias in measurement. However, due to the contextual dependency of these findings, the generalizations of the findings and their relevant implications to other contexts with test takers and raters of different backgrounds must be done with care. Since this study focused on the assessment and rating of test takers' speaking abilities by raters, further research could be done using other performance skills (e.g., writing) and in other contexts to observe whether they realize similar outcomes.

The Authors

Houman Bijani is a Ph.D. candidate in TEFL from the Science and Research Branch, Islamic Azad University, Tehran, Iran. He is also a faculty member of Zanjan Islamic Azad University. He received his M.A. in TEFL from Allameh Tabatabai University as a top student. He has published several research papers in national and international language teaching journals. His areas of interest include quantitative assessment, teacher education and language research.

Department of English Language Teaching
Science and Research Branch, Islamic Azad University, Tehran, Iran

Mobile: +989122113652
Email: houman.bijani@gmail.com

Mona Khabiri (corresponding author) is Associate Professor of Applied Linguistics at Islamic Azad University, Central Tehran Branch, and the director of the Journal of English Language Studies (JELS). She mainly teaches language testing, research methodology, seminars in TEFL issues, and teaching language skills at the graduate level, and her main areas of interest include teacher education, cooperative learning, and language testing and research. She has published papers in international and national academic journals and has presented in several national and international seminars.

Department of English Language Teaching
Central Tehran Branch, Islamic Azad University, Tehran, Iran
Email: mona.khabiri@iauctb.ac.ir

References

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

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

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.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201-219.

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

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

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., Barkhuizen, G., Knoch, U., & Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64.

Farhady, H., & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics, 29*(1), 132-141.

Fulcher, G. (1994). Some priority areas for oral language testing. *Language Testing Update, 15*(1), 39-47.

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

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

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.

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, USA.

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

Kim, Y. (2017). Score dependability in ESL oral performance assessment: A generalizability theory approach. *The Journal of Asia TEFL, 14*(3), 564-572.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study. *Language Testing, 28*(2), 179-200.

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.

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

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

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13*(4), 425-444.

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

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

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing, 25*(4), 553-581.

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

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

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.

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

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

Xi, X, & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61*(4), 1222-1255.

Zhang, B., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher ratings: competing or complementary constructs? *Language Testing, 28*(1), 31-50.