Language Testing in Asia

RESEARCH Open Access

Investigating the effect of classroom-based feedback on speaking assessment: a multifaceted Rasch analysis

Houman Bijani[1], Bahareh Hashempour[2*], Khaled Ahmed Abdel-Al Ibrahim[3,4], Salim Said Bani Orabah[5] and Tahereh Heydarnejad[6]

*Correspondence: baharehashempour1367@gmail.com

[1] Islamic Azad University, Zanjan Branch, Zanjan, Iran
[2] University of Zanjan, Zanjan, Iran
[3] Educational Psychology, College of Education, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia
[4] Sohag University, Sohag, Egypt
[5] English Language Center, University of Technology and Applied Sciences, Ibra, Oman
[6] Department of English Language, Faculty of Literature and Humanities, University of Gonabad, Gonabad, Iran

Abstract

Due to subjectivity in oral assessment, much concentration has been put on obtaining a satisfactory measure of consistency among raters. However, the process for obtaining more consistency might not result in valid decisions. One matter that is at the core of both reliability and validity in oral assessment is rater training. Recently, multifaceted Rasch measurement (MFRM) has been adopted to address the problem of rater bias and inconsistency in scoring; however, no research has incorporated the facets of test takers' ability, raters' severity, task difficulty, group expertise, scale criterion category, and test version together in a piece of research along with their two-sided impacts. Moreover, little research has investigated how long rater training effects last. Consequently, this study explored the influence of the training program and feedback by having 20 raters score the oral production produced by 300 test-takers in three phases. The results indicated that training can lead to more degrees of interrater reliability and diminished measures of severity/leniency, and biasedness. However, it will not lead the raters into total unanimity, except for making them more self-consistent. Even though rater training might result in higher internal consistency among raters, it cannot simply eradicate individual differences related to their characteristics. That is, experienced raters, due to their idiosyncratic characteristics, did not benefit as much as inexperienced ones. This study also showed that the outcome of training might not endure in long term after training; thus, it requires ongoing training throughout the rating period letting raters regain consistency.

Keywords: Bias, Interrater consistency, Intrarater consistency, Multifaceted Rasch measurement (MFRM), Rater training, Severity/leniency

Introduction

Being capable of speaking efficiently is gaining more significance in today's world; as a result, the role of teaching speaking is achieving higher prominence in second and foreign language acquisition. Therefore, speaking effectively in a second language is getting more widespread recognition as a significant skill for various life matters (Fan & Yan, 2020; Luoma, 2004). Due to the importance of speaking in the second language (SL) and foreign language (FL) contexts, speaking assessment is regarded as a vital matter. Such importance calls upon valid and reliable approaches to assessing this skill (Hughes, 2011).

One of the most significant issues related to the scoring process is the rating scale and how it is developed and used. A majority of students' performances are scored subjectively in many speaking tests by utilizing a rating scale. Scoring descriptions can then be obtained by relating the assigned number to the relevant corresponding descriptor in the scoring rubric guide (Hazen, 2020). Two related issues here are, first, the criteria selected against which the students are to be rated and, second, the number of bands or categories in the rating scale that can be justified (Moradkhani & Goodarzi, 2020).

One issue which has always been regarded as an inherent cause of evaluation error that itself might disturb the true assessment of students' speaking competence is rater variability (McNamara, 1996; Tavakoli et al., 2020). Therefore, rater effects must be considered for suitable measuring of test takers' speaking competence. A lot of research (e.g., Tavakoli et al., 2020; Theobold, 2021) on second language speaking assessment by raters has concentrated on causes of rater variation. Such variables consist of rater severity, reciprocity with other facets of the scoring setting, and inter-rater reliability (Lynch & McNamara, 1998; Rezai et al., 2022).

Without rater consistency, raters are not likely to give equal scores to a single performance; thus, severity, which is the possibility of awarding lower scores by raters, and leniency, which is the reverse aspect, are increased. This will result in the assessment being a lottery causing it to be a matter of chance that a particular test taker is scored by which rater (Ahmadi, 2019). That is, a test taker may be scored by the most lenient member of the rater group and benefit consequently or may be scored by the severest member and disadvantage as a result. Because speaking tests demand subjective assessment of this skill, much attention has been paid to achieving a satisfactory measure of consistency among raters so that scoring oral language can be done impartially and systematically (Kwon & Maeng, 2022). Nevertheless, the more emphasis is put on reliability, the less validity is obtained (Ghahderijani et al., 2021; Huang et al., 2020); in other words, emphasizing higher measures of reliability o not necessarily lead to valid measurements of speaking skills. The thing that paves the way for both a reliable and valid measurement of speaking skills is rater training.

On the contrary, McQueen and Congdon (1997) argue that although rater training is intended to maximize Interrater agreement, it does not assure the quality of assessment. Some scholars (e.g., McNamara, 1996; Weigle, 1998) have cautioned against the hazards of compulsory consistency, and as a result, have underlined individual self-consistency (intra-rater agreement) as a more fruitful goal of the training program. It is well documented that, without such training, scoring is doomed to be extremely inconsistent (Iannone et al., 2020). A fairly substantial amount of literature commencing with the research done by Huang et al. (2020) and persisting up to now with the work of Davis (2019), has been researched which establishes that training is a highly significant factor in the reliability of speaking ratings in first and second language settings accordingly. Although it is well-established that trained raters can rate students' performances reliably, there remain a number of questions about the validity of these ratings. This is due to the fact that reliable ratings do not necessarily lead to valid judgments of writing skills.

In the performance assessment, rater training has also been referred to, although from various viewpoints, especially regarding the utmost goal of achieving notable measures of consistency in scorings. Linacre (1989) specifies that unwanted error variance in scoring had better be removed or diminished as much as possible; however, there are some conceptual and theoretical obstacles to fulfilling this objective. For example, even if we train raters to assign precisely similar scores to test-takers, which is farfetched, there remain concerns regarding the interpretability of such scores.

The Multi-faceted Rasch model introduced by Linacre (2002), which can be done using the computer software FACETS, takes a different viewpoint on the issue of rater variability by considering both the factor of raters in performance-based language testing and supplying feedback to the raters based on their performance in scoring (Ahmadian et al., 2019; Lumley & McNamara, 1995). Pioneers of the Rasch technique in assessment argue that it is impossible to train raters to obtain the same degree of severity (Lunz et al., 1990). In reality, the application of the Rasch assessment rules out the requirement for bringing raters higher consistency. This is due to the fact that measures of test takers' abilities are free from those of raters' severity in assessment. However, Lumley and McNamara (1995) state that rater variation could be identified concerning the severity and random error. Thus, training and even retraining are suggested for those raters who are spotted as misfitting by the Rasch technique (Lunz et al., 1990) to provide more self-consistency (intra-rater consistency) among raters. The implication is that rater training does not intend to force raters into consistency. Consequently, as Wigglesworth (1997) suggests, the primary purpose of rater training had better be to prevent raters from implementing their subjective judgments in short intervals and as a result alter their rating approaches in long term accordingly.

Attempts in the reduction of raters' biases have produced conflicting results. In a related study, Wigglesworth (1997) found a reduction of biasedness as a result of feedback and training and that the raters were able to incorporate it in their subsequent ratings. However, more recent studies have found rather little insignificant effect (Elder et al., 2007; John Bernardin et al., 2016; Rosales Sánchez et al., 2019). Wigglesworth (1997) investigated bias in the context of rater training to evaluate both live and tape-based oral tests. She observed different behaviors and significant variations in how the rates responded to various test criteria based on the modality of the interview. Some raters were severer on fluency or vocabulary, while some others rated them more leniently. Also, they were different on account of their severity estimates for different task types. However, it seemed that raters were able to incorporate the feedback they received in their subsequent ratings since their level of biasedness was reduced to a considerable extent compared to the previous ratings. However, in Wigglesworth's study raters were first given the feedback and then attended a group rater training session; therefore, it is not clear whether the changes in the rating behavior are due to individual feedback or both the bias reports and rater training session. Fan and Yan (2020) investigated the consistency of raters' severity/leniency over several grading intervals by using MFRM. The outcomes of data analysis demonstrated significant instability in two of the three scoring periods ranging from one to four. In another study, Lumley and McNamara (1995) investigated three sets of grading of a spoken English test in 20 months. The findings of the interactional effects of time and rater facets represented a significant change in rater's severity. Studies conducted by Brown (2005) revealed that trained, native, or advanced speakers of the target language raters, score test takers with the same degree of severity and consistency as other raters do.

The use of MFRM in bias analysis has got several implications for performance assessment. First, MFRM helps researchers study the rater facet concerning their facet of interest by keeping the other facets constant and neutral (Lunz et al., 1990). Second, it can help researchers in administering rater training programs. Research has shown that rater consistency and rating validity can be increased through training (McQueen & Congdon, 1997). Third, MFRM can help reduce self-inconsistency and increase intrarater reliability, which increases the fairness of tests specifically in placement and summative evaluation tests (Davis, 2019). Tavakoli et al. (2020) investigated the rating of 40 essays written by Japanese students by employing 40 native English speakers. Each rater scored all the 40 essays on a six-point analytic rating scale of five categories. The results showed that some raters scored higher ability test-takers more severely and lower ability ones more leniently than expected. Brown (2005) studied rater bias in a face-to-face oral test of Japanese EFL learners. The results of MFRM showed significant bias in scoring criteria but no significant bias in task fulfillment.

Nevertheless, much of the research done up to now has explored the use of FACETS on just a couple of facets. For instance, research on the rater's severity or leniency on test-takers (Lynch & McNamara, 1998), task types (Wigglesworth, 1997), and specific rating time (Lumley & McNamara, 1995; Vadivel & Beena, 2019). However, no research has incorporated the facets of test takers' ability including facets of test takers' ability, raters' severity, task difficulty, group expertise, scale criterion category, and test version so far all in one piece of research together with their two-sided impacts.

Even though earlier research on rater variation has emphasized achieving higher measures of raters' consistency as the ultimate aim of rater training (Bijani & Fahim, 2011; Lumley & McNamara, 1995; McNamara, 1996; Vadivel et al., 2021), rater variability can be still traced following training not only for rater severity but also for internal consistency. Also, the dynamic and unpredictable nature of oral interaction questions the reliability of the measure of oral competence. This unpredictability will also affect test validity. In other words, test takers may receive different scores on different occasions from different raters. There is a considerable amount of research exploring the discourse of oral language interviews (e.g., Brown, 2005); however, little research has ever investigated the variation among raters.

Although it is verified that rater training has a significant role in persuading higher consistency among raters in terms of their rating behaviors, there is still a paucity of information about how training functions to provide higher measures of consistency among raters. Even if several rater training impacts have been specified, there are still few studies stipulating such impacts (Brown, 2005; Liu et al., 2021). In addition to that, little research has explored the duration of rater training effects (Bijani, 2010). There are studies exploring the effectiveness of the training program in short periods, but few studies have investigated its effectiveness after a long period following training since Lumley and McNamara (1995) suggested that the outcomes of training might not endure in long terms following training and that raters may change over time. Thus, a need for renewed training is worth investigating.

The results of this study can provide fruitful information to teachers who are doing their pre-service teacher education programs or teachers who are already doing their in-service education programs. Since teachers, might assess students' speaking performances for a variety of reasons, the provision of opportunities to practice rating in a way that is accompanied by individual rating feedback can help raters improve their rating ability. Moreover, the results of this study could be used for raters with various degrees of rating proficiency—inexperienced and experienced ones. Also, the results of this study can demonstrate characteristics to be used in rater training and teacher education.

Therefore, this study focused on raters' severity, bias and interaction measures, and internal consistency considering their interaction of the six different facets used in the study including test takers' ability, rater severity, raters' group expertise, task difficulty, test version, and rating scale criteria using a quantitative approach. Each rater's rating behavior was primarily analyzed so that it would provide feedback to them accordingly. Then, an investigation of the scoring behaviors of the two groups of raters (experienced and inexperienced raters) was followed. Besides, this study investigated the enhancement of rating ability through lapse of time in both rater groups. Also, the two groups of raters were compared with each other in each rating session. Therefore, the following search questions can be formed:

1. How much of test takers' total score variance can be accounted for in each facet (test takers' ability, raters' severity, task difficulty, group expertise, scale criterion category, and test version)?
2. To what extent was the provided feedback successful following the training program regarding severity, bias, and consistency measures?

Methodology

Design

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

Participants

As many as 300 adult Iranian students of English as a Foreign Language (EFL), consisting of 150 males and 150 females, between the ages of 17 and 44 took part in this research as test-takers. The participants were chosen from a pool of Intermediate, High-intermediate, and Advanced stages of learning English at the Iran Language Institute (ILI). The reason for opting for the students from the aforementioned levels of proficiency was due to the fact that they had already acquired the required fundamental principles of academic oral performance.

It was mentioned that the test takers were selected from three various English proficiency levels at the ILI; however, considering the sole educational level could not be a valid criterion for classifying learners into different proficiency levels. Thus, to make sure that the test-takers taking part in this study were not at the same level of language proficiency, a TOEFL test was given to make sure whether there was a significant difference between them or not. In order to make sure whether there is a significant mean difference among the scores of the test takers of the three groups, an ANOVA was run. Table 1 demonstrates the ANOVA statistical analysis of the TOEFL scores of the three groups of test-takers.

The outcome shows that there is a significant difference with respect to takers' general language proficiency (TOEFL score) among the test takers.

As many as 20 Iranian EFL teachers, consisting of 10 males and 10 females, between the ages of 24 and 58 took part in this research as raters. The raters were Bachelor's and Master's holders in English language-related majors, working in various public and private academic centers. As one of the prerequisites of this study, the raters had to be separated into groups of experienced and inexperienced ones in order to explore their similarities and differences and to investigate which group might outperform the other one. In addition to that, to keep the data provided by the raters confidential, their names and identities were anonymized by attributing them each a score from 1 to 10.

The raters were provided with a background questionnaire, adapted from McNamara and Lumley (1997), with the help of which information included (1) *demographic information*, (2) *rating experience*, (3) *teaching experience*, (4) *rater training*, and (5) *relevant courses passed* would be obtained. The obtained data are summarized in Table 2.

Thus, the raters were classified into two expertise groups based on their experiences stated above.

A. Raters with no or fewer than two years of experience in rating and undertaking rater training, plus no or fewer than five years of experience in English language teaching and managed to pass fewer than the four core courses relevant to English language teaching. From now on these raters are referred to as *new* raters.
B. Raters with two and more years of experience in rating and undertaking rater training, plus five and more years of experience in English language teaching and managed to pass all the four core courses relevant to English language teaching as well as a minimum of two other selective courses. From now on these raters are referred to as *old* raters.

Table 1 ANOVA table for the TOEFL scores of the three groups of test takers

	Sum of squares	*df*	Mean square	*F*	Sig.
Between groups	23424.620	2	11712.310	2197.362	0.000
Within groups	1583.060	297	5.330		
Total	25007.680	299			

Table 2 Criteria for rating expertise

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

Table 3 Rater background characteristics

Raters	*N*	Male	Female	Mean age	Rating experience	Teaching experience	Rater training	Relevant courses passed
NEW	10	5	5	41.2	0.8	3.7	0.3	2.4
OLD	10	5	5	31.7	3.4	14.2	4.1	4.7

A more important reason for choosing these groups of expertise is to investigate any differences between experienced and inexperienced raters in terms of how they approach the task of oral assessment and how they are affected by the rating process. It is noteworthy to indicate that in order to eliminate the rater expectancy effect, the raters and rater groups were not informed of the existence of two various groups and any similarities and differences between the two. Table 3 displays the summary characteristics of the raters participating in the study.

It is noteworthy to indicate that all the participants were informed in advance that they were participating in a research study and the researchers obtained their consent orally that the outcomes of this research would be used to make publications, yet their identities would be kept anonymous.

Instruments

The present study aimed to use the Community English Program (CEP) test to evaluate test takers' speaking ability in different settings. The goal of the speaking test is to evaluate to what extent the speakers of a second language can produce meaningful, coherent, and contextually appropriate responses to the following five tasks.

Task 1 (*description task*) is an independent-skill task that displays the personal experience of test-takers to answer without input provision (Bachman & Palmer, 1996). Moreover, task

3 (*summarizing task*) and 4 (*role-play task*) display test takers' listening ability in responding orally to any given input. In other words, the response contents are given to the test takers via short and long listening. For tasks 2 (*narration task*) and 5 (*exposition task*), the test takers are needed to give a response to pictorial prompts consisting of a series of photos, graphs, figures, and tables.

The aforementioned tasks were implemented via two delivery methods: (1) direct and (2) semi-direct. The former is aimed to use for an individual face-to-face method; however, the semi-direct test is mainly aimed for use in a language laboratory context.

As one of the requirements of this study to evaluate the influence of using a scoring rubric on the validity and reliability of assessing test takers' oral ability, this study aimed to employ an analytic rating scale. The purpose of using an analytic rating scale was to assess test takers' oral performance to determine the extent to which it evaluates the oral proficiency of test-takers more validly and reliably. For either version of the test, all the test takers' task performances were evaluated by the use of the ETS (2001) analytic rating scale. In ETS (2001) rating scale, evaluation is done based on *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.

The reliability of the test was estimated. According to Table 4, the reliability of the questionnaire, in whole including 20 items, was $\alpha \geq 87.7\%$ which is according to Cohen's table of effect size considered much larger than typical.

Also, to ascertain the validity of the test, a confirmatory factor analysis (CFA) was run, and the obtained model fit reflecting the result of CFA displayed NFI (normal fit index) = 0.91, CFI (comparative fit index) = 0.92, TLI (Tucker-Lewis index) = 0.95, SRMR (standardized root mean square residual) = 0.06, RMSEA (root mean square error of approximation) = 0.042. All the obtained indices indicate the goodness of the model and confirm the validity of the questionnaire.

Procedure

Pre-training phase

The 300 students were randomly selected to take a sample TOEFL (iBT) test including listening, structure, and reading comprehension to make sure that they are not at the same level of language proficiency and that there is a significant difference between the three groups. Meanwhile, the raters were awarded the background questionnaire before running the test tasks and collecting data. As indicated before, this was intended to separate the raters into the two groups of experienced and inexperienced ones.

Having made sure that the three groups of test-takers are at various levels of language proficiency and identified the raters' background information and their level of expertise and classified them as inexperienced raters and experienced ones, the speaking test started. It is worthy to indicate that the 300 test-takers who took part in this research were separated into three groups where each would take part in a stage of this research namely (pre-/

Table 4 Reliability statistics of the CEP test

Cronbach's alpha	*N* of items
.824	300

immediate post-/delayed post-training). Half of the members of each group would also participate in the direct and the other half in the semi-direct version of the speaking test. The reason why all the raters did not take part in both versions of the oral test was owing to the impact of each version that would most possibly influence their performance in the other test version. Such an action would familiarize the raters with the type of questions appearing in either version and would thus negatively influence the validity of the research. The raters were then given a week to submit their ratings, based on the six-band analytic rating scale, to the researcher.

Rater training procedure

Once the pre-training phase was over, the raters took part in a training or norming session during which they got familiar with the oral tasks and the rating scales. The training program was done by the first author of this article who is an authorized IELTS instructor and a Ph.D. holder in the field of Teaching English as a Foreign Language (TEFL). He trained the raters in two sessions and also evaluated all test takers' performance in the three phases of the study to serve as benchmarks for raters' ratings and further data analysis.

They also had the opportunity to practice the instructed materials provided with a number of sample responses collected from some similar proficiency-level students other than the ones participating in this study as test-takers. The researcher gave each rater information about the scoring process as the objective of the training program was to make raters with various degrees of expertise familiar with significant aspects of scoring while they score each student's speech production.

In the meantime, the responses which were previously recorded were played for the raters as they were monitored and provided with direct guidance from the trainer. The raters were also encouraged to form panel discussions and share their justifications and reasons behind the scores they decided to assign while giving reference to the scoring rubric.

The trainer also provided individual feedback for each rater regarding their previous ratings during the pre-training phase. This is what Wallace (1991) stresses in rater training programs. He believes that what helps acquired knowledge to get internalized is through reflection not merely by repeated practice. This will further provide the raters with a chance to reflect upon their scoring behavior. Since each rater possesses a different rating ability and rating behavior, each rater needed to be provided with feedback individually.

Immediate post-training phase

Immediately following the rater training program, when the raters got the required skill in rating speaking ability, the speaking tasks (description, summarizing, role-play, narration, and exposition) of both versions of the test (direct and semi-direct) were administered one by one. As it was mentioned before in the pre-training data collection procedure, the second third of the test takers (including 100 students) were tested from whom the data were elicited. It is again stressed that the oral tasks were assessed using the ETS rating scale. The selection of 100 oral performance data in whole ([100 × 5] semi-direct + [100 × 5] direct = 1000 in general) of both methods at this stage was done randomly for each rater. Randomization was done to counteract the influence of sequencing the performances on the raters' behaviors so that they could not remember how many data at a particular score were rated by them.

Delayed post-training phase

Exactly 2 months (as suggested by McNamara, 1996) after the immediate post-training data collection, the fourth phase of the data collection procedure was administered. In this phase, the last third of the test takers (including 100 students) were tested from whom data were elicited. The raters were provided with the collected data to rate based on the knowledge they had already gained during the rater training program two months before. The aim was to observe the delayed impact of the training program on raters and also the degree of inter-rater reliability. The expectation was that raters were still consistent in rating.

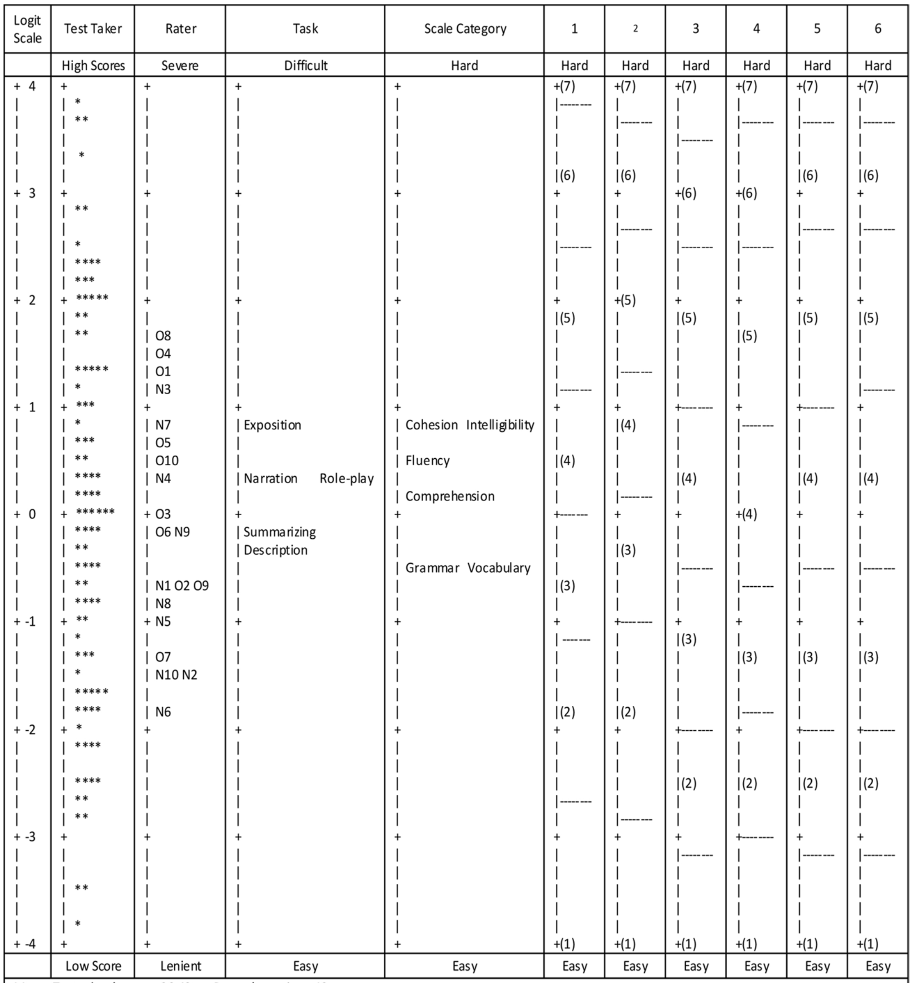
Data analysis

Quantitative data (i.e., raters' scores based on an analytic rating scale) were gathered and analyzed with MFRM during three scoring sessions. In order to compensate for the impact of test methods and rating factors, MFRM has been widely used in second/foreign language performance assessments. Through estimating the probabilities of patterns of responses, MFRM estimates the variability associated with the factors such as test-takers, raters, rater groups, tasks, and rating scales involved during performance assessment procedures (McNamara, 1996). MFRM also provides information on rater characteristics, specifically severity, consistency, randomness, and inter-rater reliability (McNamara, 1996). Consequently, studies on rater variability in the speaking and writing scoring process have often used MFRM in their analyses (McNamara, 1996; Weigle, 1998). In the present study, MFRM was performed using the FACETS program after each rating stage or phase of the study to examine both individual rater and rater group scoring patterns. A six-facet model will be used including the facets of the test taker (test takers' ability), rater (raters' severity/leniency), rater group (experienced/inexperienced), task (the tasks used in the study), rating criterion (categories of the rating scale for the analytic scale), and test method (direct/semi-direct). Therefore, a six-facet partial credit model will be employed.

The patterns of the awarded scores of the two groups of raters (new and old) were investigated each time they rated test takers' oral performances by the use of an analytic rating scale. The quantitative data were compared (1) across the two groups of raters to explore the raters' capability cross-sectionally at each scoring stage, and (2) within each rater group to study the improvement of the raters' ability.

Results

Having analyzed the data during the pre-training phase, the FACETS variable map representing all the facets was obtained. In the FACETS variable map, presented in Fig. 1, the facets are placed on a common logit scale that facilitates interpretation and comparison across and within the facets in one report. The figure plots test takers' ability, raters' severity, task difficulty, scale criterion difficulty, test version difficulty, and group expertise. According to McNamara (1996), the logit scale is a measurement scale that



Mean: Test takers' scores: 20.43 Raters' severity: -.12
SD: Test takers' scores: 3.07 Raters' severity: 1.12
Rater' separation index: 3.69 Reliability: .92
Test takers' separation index: 7.50 Reliability: .89

Fig. 1 FACETS variable map (pre-training)

expresses the probabilities of test takers' responses in various conditions of measurement. It also contains the means and standard deviations of the distributions of estimates for test-takers, raters, and tasks at the bottom.

The *first column (logit scale)* in the map depicts the logit scale. It acts as a fixed reference frame for all the facets. It is a true interval scale that has got equal distances between the intervals (Prieto & Nieto, 2019). Here, the scale ranges from 4.0 to – 4.0 logits.

The *second column (Test Taker)* displays estimates of test takers' proficiency. Each star displays a singlet test taker. Higher scoring (more competent) test takers are at the top of the column whereas lower scoring (less competent) ones are at the bottom. Here, the range of the test takers proficiency ranges from 3.81 to – 3.69 logits; thus making a spread of 7.50 with respect to test takers' ability. It is worthwhile to specify that no test taker was identified as misfitting; thus, none of them was excluded from data analysis during the pre-training phase of this research.

The *third column (rater)* displays raters concerning their severity or leniency estimates in scoring test takers' oral proficiency. Since there was more than one rater scoring each test taker's performance, raters' severity or leniency scoring patterns can be estimated. This will give us raters' severity indices. In this column, each star displays one rater. Severer raters appear at the top of the column, whereas more lenient ones at the bottom. At the pre-training, rater OLD8 (severity measure 1.72) was the severest rater and rater NEW6 (severity measure – 1.97) was found to be the most lenient rater. Besides, in this phase, OLD raters, on average, were rather severer than NEW raters who tended to be more lenient than the OLD ones. Here, raters' severity estimate ranges from 1.72 to – 1.97 logits which makes the distribution of rater severity measures (logit range = 3.69) which is much narrower than the distribution of the test takers' proficiency measures (logit range = 7.50) in which the highest and lowest proficiency logit measures were 3.81 and – 3.69 respectively. This demonstrates that the effect of individual differences on behalf of raters on test-takers was relatively small. Raters, as shown in the figure, seem to have spread equally above and below the 0.00 logits.

The *fourth column (task)* displays the oral tasks used in this study in terms of their difficulty estimates. The tasks appearing at the top of the column are harder for the test takers to implement than the ones at the bottom. Here, the *Exposition task* (logit value = 0.82) was harder for the test takers than the other tasks, while the *Description task* (logit value = – 0.37) was the least difficult one; therefore, making a spread of 1.19 logit range variation. This column has the lowest variation in which all the elements are gathered around the mean.

The *fifth column* (scale category) displays the severity of scoring the rating scale categories. The most severely rated scale category appears at the top and the least severely rated scale category appears at the bottom. Here, *Cohesion* was measured to be the most severely scored category (logit value = 0.79) for raters to use whereas *Grammar* was the least severely scored one (logit value = – 0.46).

Columns 6 to 11 (rating scale categories) display the six-point rating scale categories employed by the raters to evaluate the test takers' oral performances. The horizontal lines across the columns are the categories threshold measures that specify the points at which the probability of achieving the next rating (score) starts. The figure shows that each score level was used although there was less frequency at the extreme points. Here, the test takers with the proficiency measure of between – 1.0 and + 1.0 logits were likely to get ratings of 3 to 4 in *Cohesion*. Similarly, the test takers at the logit proficiency of 2.0 logits had a relatively high probability of receiving a 5 from a rater at the severity level of 2.0 in *Intelligibility*.

RQ1: How much of test takers' total score variance can be accounted for in each facet?

A FACETS program enables us to determine how much each score variance is attributed to which of the facets employed. Accordingly, one more data analysis was done to measure to what extent the total score variance is associated with each of the facets identified in this study. Table 5 shows the percentage of total score variance associated with each of the facets used in the study prior to the training program. The information provided in the table shows that the greatest percentage of the total

Table 5 Effect of each facet on total score variance (pre-training)

No.	Facets identified in the study	Percentage effect on total score variance
1	Test taker ability	44.82
2	Rater severity	26.13
3	Group expertise	14.67
4	Test version	6.58
5	Task difficulty	4.74
6	Scale categories	3.06
		100

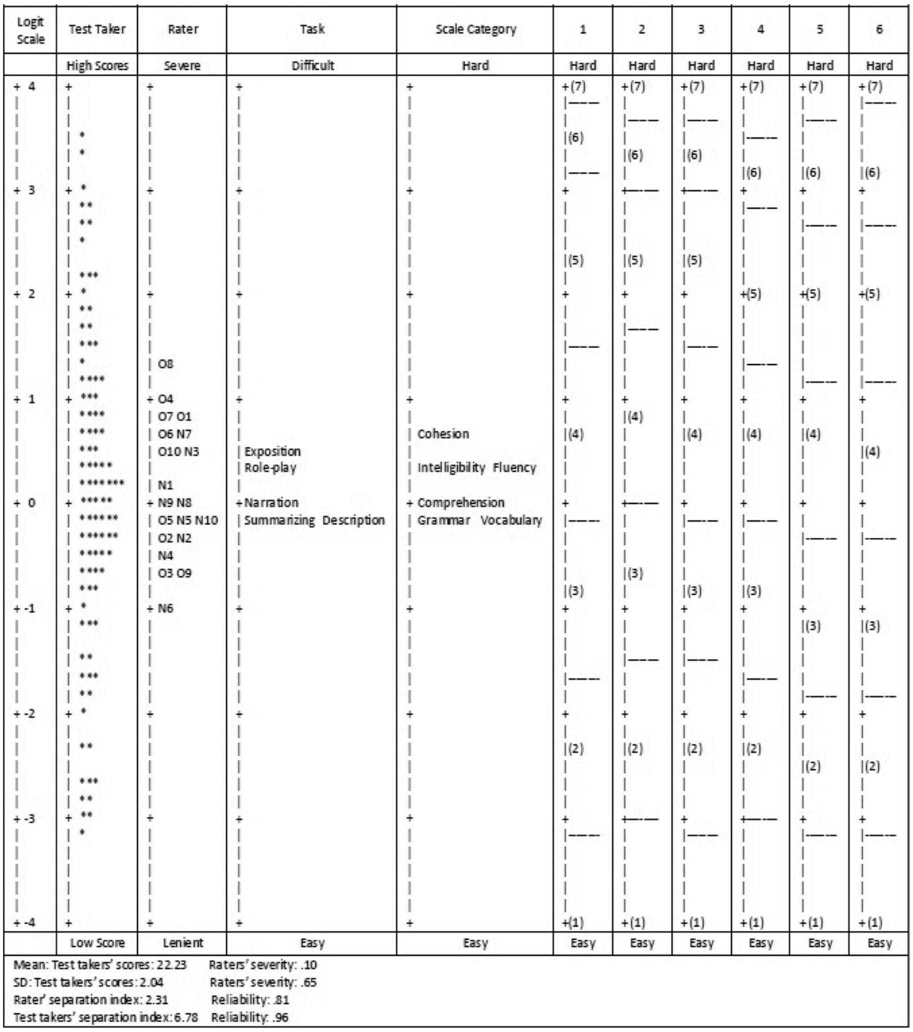
variance (44.82%) is related to the test takers' ability differences however the remaining variance (55.18%) is related to other facets including rater's severity, group expertise, test version, task difficulty, and scale categories.

The rather high percentage of total score variance, other than that of test takers' capability at the pre-training phase calls up the caution to be taken about the effect of unsystematicity of rating and the existence of undesirable facets influencing the final obtained score. Furthermore, it shows that the rater's facet entails a significant extent of total test variance (26.13) which indicates that there is a likelihood of inconsistency and disagreement between raters and their judgments proving that a number of raters are relatively severer or more lenient towards the test takers than the other raters. This finding represents that the test-takers will be scored differently depending on the rater. The rather small effect of other facets including test version, task difficulty, and scale categories shows that there is a slight bilateral and multilateral interactional effect of the facets involved in test variability; thus, proving the neutralizing effect of test variability through the combination of other test facets.

Having analyzed the data at the immediate post-training phase, the FACETS variable map representing all the facets and briefly stating the main information about each one was obtained. The FACETS variable map, displayed in Fig. 2, plots test takers' ability, raters' severity, task difficulty, scale criterion difficulty, test version difficulty, and group expertise.

The *second column (test taker)* displays estimates of test takers' proficiency. Here, the range of the test takers' proficiency ranges from 3.62 to – 3.16 logits, with a spread of 6.78 logit value. The reduction of test takers' proficiency logit from 7.50 (before training) to 6.78 (after training) shows that they were rated more similarly with regard to severity/leniency indices. This reflects that the test takers have been more clustered around the mean concerning raters' scoring of their oral proficiency level.

The *third column (rater)* displays raters about severity or leniency estimates in rating test takers' oral proficiency. Here, raters' severity estimate ranges from 1.26 to – 1.05 logits which makes the distribution of rater severity measures (logit range = 2.31) which is again a lot narrower than (almost one-third) the distribution of the test takers' proficiency measures (logit range = 6.78) in which the highest and lowest proficiency logit measures were 3.62 and – 3.16 respectively. This demonstrates that the effect of individual differences on behalf of raters on test-takers was relatively small.



```
Logit Test Taker Rater        Task                      Scale Category            1       2       3       4       5       6
Scale
      High Scores Severe       Difficult                 Hard                      Hard    Hard    Hard    Hard    Hard    Hard
+ 4   +          +            +                         +                         +(7)    +(7)    +(7)    +(7)    +(7)    +(7)
|     |          |            |                         |                         |---    |       |       |       |       |---
|     |          |            |                         |                         |       |---    |---    |       |---    |
|     | *        |            |                         |                         |(6)    |       |       |---    |       |
|     | *        |            |                         |                         |       |(6)    |(6)    |       |       |
|     |          |            |                         |                         |---    |       |       |(6)    |(6)    |(6)
+ 3   + *        +            +                         +                         +       +---    +---    +       +       +
|     | **       |            |                         |                         |       |       |       |---    |       |
|     | **       |            |                         |                         |       |       |       |       |---    |---
|     | *        |            |                         |                         |       |       |       |       |       |
|     |          |            |                         |                         |(5)    |(5)    |(5)    |       |       |
|     | ***      |            |                         |                         |       |       |       |       |       |
+ 2   + *        +            +                         +                         +       +       +       +(5)    +(5)    +(5)
|     | **       |            |                         |                         |       |       |       |       |       |
|     | **       |            |                         |                         |       |---    |       |       |       |
|     | ***      |            |                         |                         |---    |       |---    |       |       |
|     | *        | O8         |                         |                         |       |       |       |---    |       |
|     | ****     |            |                         |                         |       |       |       |       |---    |---
+ 1   + ***      + O4         +                         +                         +       +       +       +       +       +
|     | ****     | O7 O1      |                         |                         |       |(4)    |       |       |       |
|     | ****     | O6 N7      |                         | Cohesion                |(4)    |       |(4)    |(4)    |(4)    |
|     | ***      | O10 N3     | Exposition              |                         |       |       |       |       |       |(4)
|     | *****    |            | Role-play               | Intelligibility Fluency |       |       |       |       |       |
|     | *******  | N1         |                         |                         |       |       |       |       |       |
+ 0   + *****    + N9 N8      +Narration                + Comprehension           +       +---    +       +       +       +
|     | ******   | O5 N5 N10  | Summarizing Description | Grammar Vocabulary      |---    |       |---    |---    |       |
|     | ******   | O2 N2      |                         |                         |       |       |       |       |---    |---
|     | *****    | N4         |                         |                         |       |       |       |       |       |
|     | ****     | O3 O9      |                         |                         |       |(3)    |       |       |       |
|     | ***      |            |                         |                         |(3)    |       |(3)    |(3)    |       |
+ -1  + *        + N6         +                         +                         +       +       +       +       +       +
|     | ***      |            |                         |                         |       |       |       |       |(3)    |(3)
|     |          |            |                         |                         |       |       |       |       |       |
|     | **       |            |                         |                         |       |---    |---    |       |       |
|     | ***      |            |                         |                         |---    |       |       |---    |       |
|     | **       |            |                         |                         |       |       |       |       |---    |---
+ -2  + *        +            +                         +                         +       +       +       +       +       +
|     |          |            |                         |                         |       |       |       |       |       |
|     | **       |            |                         |                         |(2)    |(2)    |(2)    |(2)    |       |
|     |          |            |                         |                         |       |       |       |       |(2)    |(2)
|     | ***      |            |                         |                         |       |       |       |       |       |
|     | **       |            |                         |                         |       |       |       |       |       |
+ -3  + **       +            +                         +                         +       +---    +       +---    +       +
|     | *        |            |                         |                         |---    |       |---    |       |---    |---
|     |          |            |                         |                         |       |       |       |       |       |
|     |          |            |                         |                         |       |       |       |       |       |
|     |          |            |                         |                         |       |       |       |       |       |
|     |          |            |                         |                         |       |       |       |       |       |
+ -4  +          +            +                         +                         +(1)    +(1)    +(1)    +(1)    +(1)    +(1)
      Low Score  Lenient      Easy                      Easy                      Easy    Easy    Easy    Easy    Easy    Easy
Mean: Test takers' scores: 22.23          Raters' severity: .10
SD: Test takers' scores: 2.04             Raters' severity: .65
Rater' separation index: 2.31             Reliability: .81
Test takers' separation index: 6.78       Reliability: .96
```

Fig. 2 FACETS variable map (immediate post-training)

Likewise, in the pre-training phase, raters, as shown in the figure, seem to have spread equally above and below the 0.00 logits. Besides, the significant reduction of raters' severity measure distribution from 3.69 in the pre-training phase to 2.31 in the immediate post-training phase displays the efficiency of the training program in bringing raters closer to one another concerning severity/leniency indices. In other words, they rated more similarly concerning severity/leniency after the training program.

The *fourth column (task)* displays the oral tasks used in this study in terms of their difficulty estimates. Here, the *Exposition task* (logit value = 0.61) was harder for the test takers than the other tasks while the *Description task* (logit value = – 0.14) was the least difficult one; therefore, making a spread of 0.75 logit range variation. The reduction of logit range, compared to the pre-training phase, indicates that the tasks were rated with less severity and leniency. This column has the lowest variation in which all the elements are gathered around the mean.

The *fifth column (scale category)* displays the rating scale category severity in scoring. Here, *Cohesion* was measured to be the most severe category (logit value = 0.58) for raters to use whereas *Grammar* was the least severe one (logit value = –0.17).

Similar to the pre-training phase, the total score variance attributable to each facet was calculated to measure the effect of each facet on total score variance immediately following the training program. Table 6 displays the percentage of total score variance associated with each of the facets used in the study at the immediate post-training phase. The information provided in the table shows that the greatest percentage of the total variance (67.12%) is related to the test takers' ability differences however the remaining variance (32.88%) is related to other facets including rater's severity, group expertise, test version, task difficulty, and scale categories.

The considerable increase of total score variance percentage attributed to test takers' ability and reduction of variance percentage attributed to other facets indicates the significant increase of systematicity and consistency in scoring following the training program. In other words, the training program was quite effective in the reduction of undesirable facets and unsystematicity of scoring influencing total score variance in the immediate post-training phase. The scoring procedure moved towards the establishment of consistency in scoring in a way that a majority of score variance was associated to test takers' performance ability differences.

Having analyzed the data at the delayed post-training phase of this research, the FACETS variable map representing all the facets was obtained. The FACETS variable map, displayed in Fig. 3, plots test takers' ability, raters' severity, task difficulty, scale criterion difficulty, test version difficulty, and group expertise.

The *second column (test taker)* displays estimates of test takers' proficiency. Here, the range of the test-taker's proficiency ranges from 3.70 to – 3.53 logits, with a logit distribution of 7.23.

The *third column (rater)* displays raters concerning their severity or leniency estimates in rating test takers' oral proficiency. Here, raters' severity estimate ranges from 1.28 to – 1.26 logits which makes the distribution of rater severity measures (logit range = 2.54) which is again a lot narrower than (almost one-third) the distribution of the test takers' proficiency measures (logit range = 7.23) in which the highest and lowest proficiency logit measures were 3.70 and – 3.53 respectively. This demonstrates that the effect of individual differences on behalf of raters on test-takers was relatively small. Similar to

Table 6 Effect of each facet on total score variance (immediate post-training)

No.	Facets identified in the study	Percentage effect on total score variance
1	Test taker ability	67.12
2	Rater severity	19.31
3	Group expertise	6.77
4	Test version	3.16
5	Task difficulty	2.12
6	Scale categories	1.52
		100

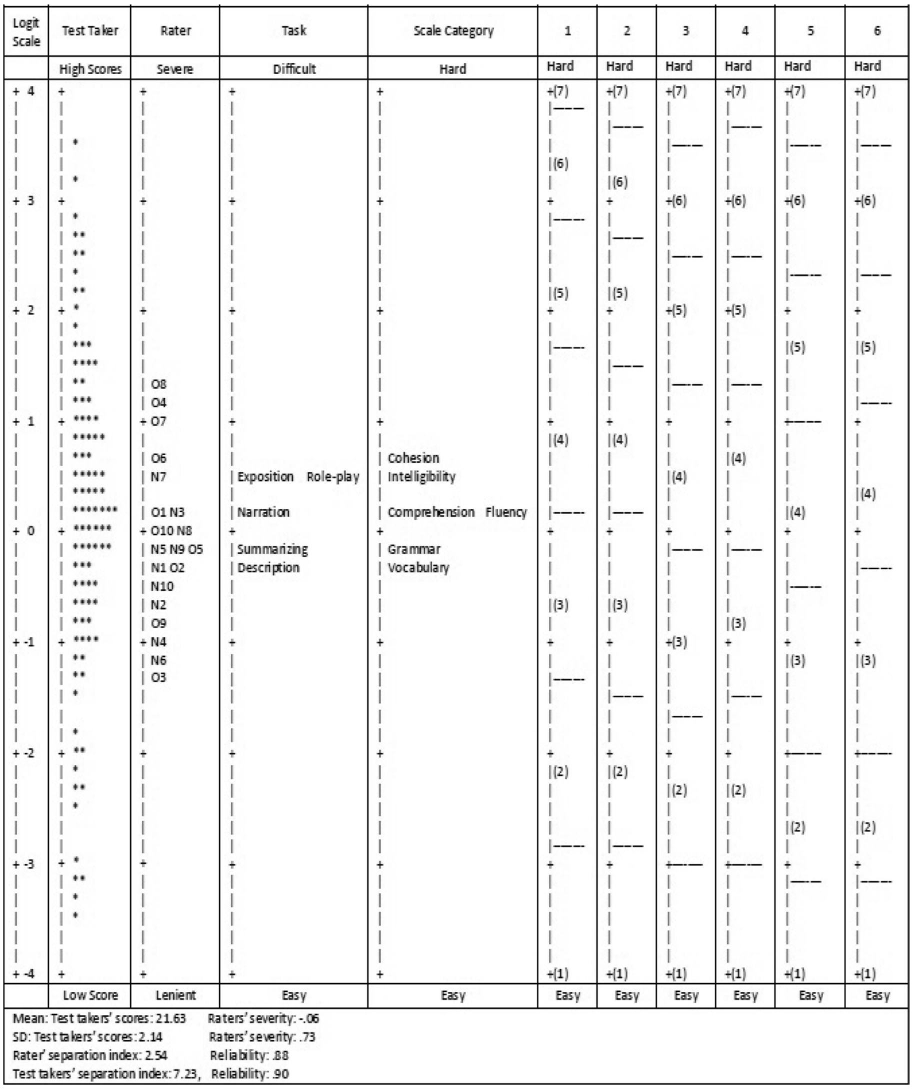


Fig. 3 FACETS variable map (delayed post-training)

the previous two phases of the study, raters, as shown in the figure, seems to have spread equally above and below the 0.00 logits. Through comparing the measures of severity distribution, raters were still closer to one another in the delayed post-training phase (2.54 logits) regarding severity/leniency measure compared to the pre-training phase (3.69 logits) which shows the rather long-lasting effectiveness of the training program. However, the increase in severity logit measure compared to the immediate post-training phase (2.31 logits) reflects the raters' tendency in moving gradually to their way of rating which implied a need for ongoing training programs in specific intervals.

The *fourth column (task)* displays the oral tasks used in this study regarding their difficulty estimates. Here, the *Exposition task* (logit value = 0.66) was harder for the test takers than the other tasks while the *Description task* (logit value = – 0.24) was the least difficult one. This column has the lowest variation in which all the elements are gathered around the mean.

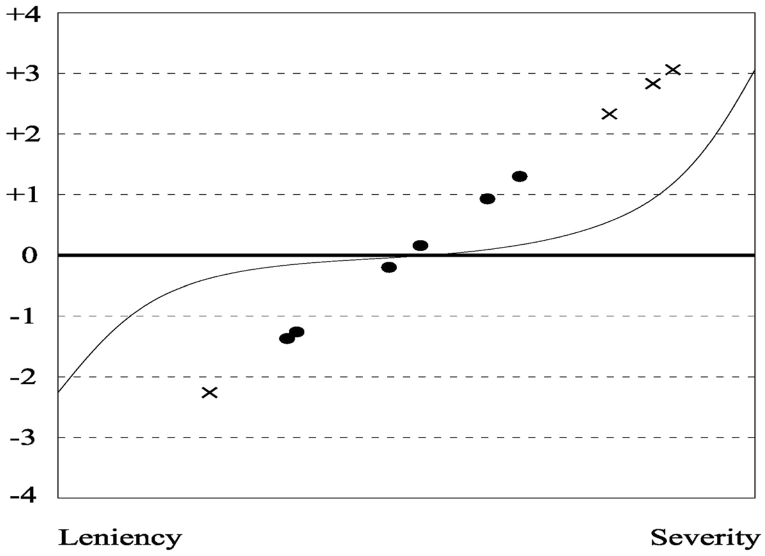


Fig. 4 Old raters' bias interaction (pre-training)

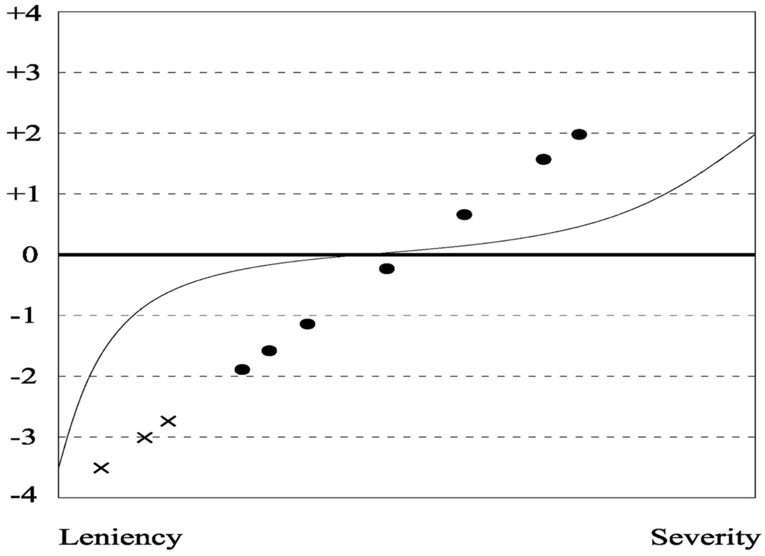


Fig. 5 New raters' bias interaction (pre-training)

The *fifth column (scale category)* displays the rating scale category severity of scoring. The most severely scored scale category was at the top and the least severely scored scale category was at the bottom. Here, *Cohesion* was measured to be the most severely scored category (logit value = 0.62) for raters to use whereas *Vocabulary* was the least severely scored one (logit value = – 0.24).

Figures 4, 5, 6, 7, 8, and 9 graphically plot the raters' bias interaction curve to the test-takers in *Z*-scores for new and old raters at the three phases of the study. The graphs display all rater biases be they significant or not. In each plot, the curved line displays the raters' severity logit. The symbols ● show *z*-scores that indicate non-significant bias, and the ✖ symbols indicate significant bias.

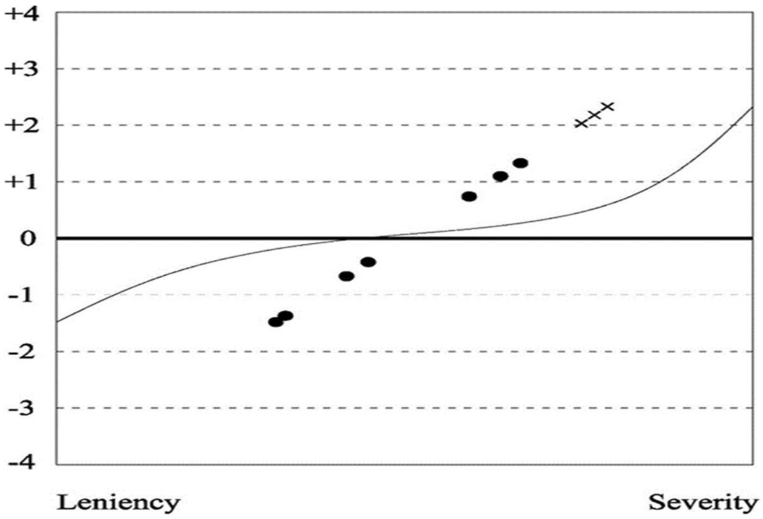


Fig. 6 Old raters' bias interaction (immediate post-training)

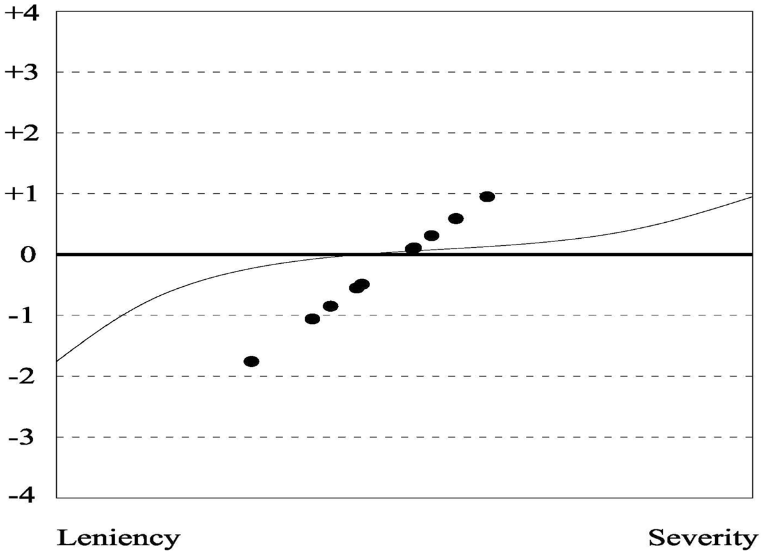


Fig. 7 New raters' bias interaction (immediate post-training)

Pre-training: there were 3 significant biases for NEW raters which were identified as significantly lenient. For old raters, the data showed 4 significant biases among which 3 were identified as significantly severe and 1 lenient.

Immediate post-training: there were 3 significant biases for OLD raters which were identified as significantly severe. No NEW raters were spotted to have a significant bias in the immediate post-training phase of the study.

Delayed post-training: there was 1 significant bias for NEW raters who were identified as significantly lenient; however, the leniency was slightly below the acceptable range which could be ignored, too. For OLD raters, the data showed 4 significant

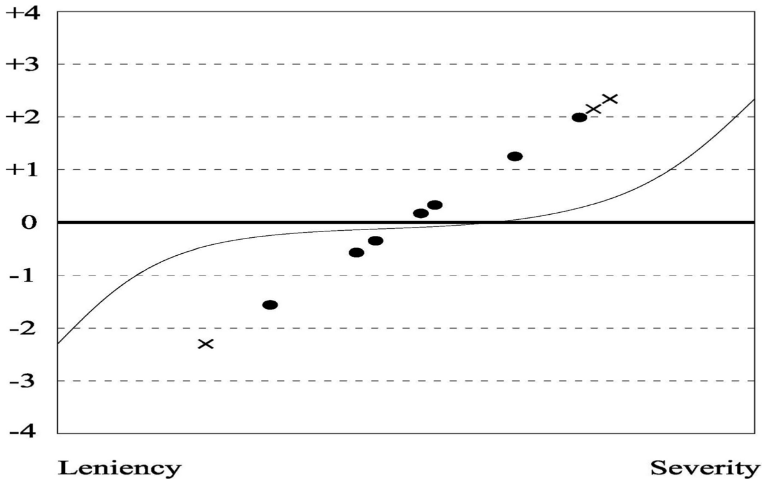


Fig. 8 Old raters' bias interaction (delayed post-training)

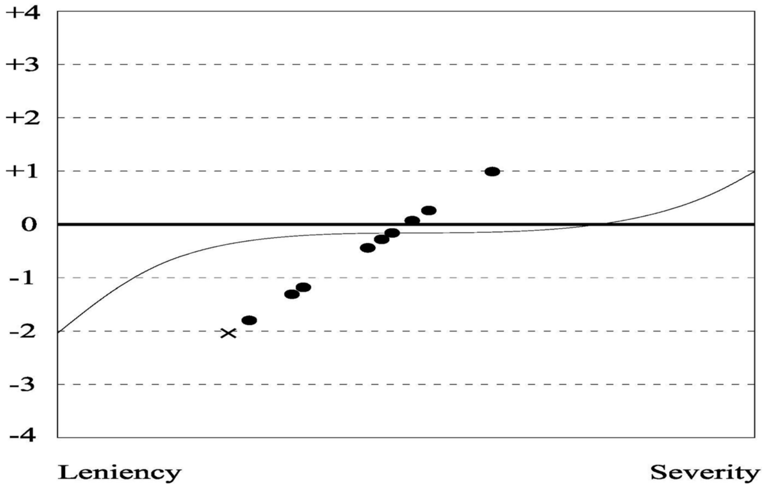
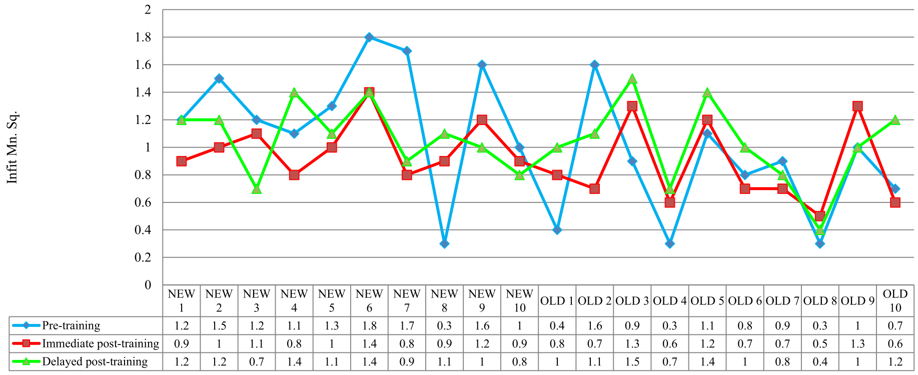


Fig. 9 New raters' bias interaction (delayed post-training)

biases among which 3 were identified as significantly severe and 1 lenient. One rater was on the borderline of severity measure.

Additionally, in order to graphically represent the raters' consistency measures throughout the three phases of the study, the raters' infit mean square values were employed. As indicated before, the infit mean square that ranges between 0.6 and 1.4 is considered the acceptable range (Wright & Linacre, 1994). The following figure (Fig. 10) plots graphically the change of raters' consistency in rating using infit mean square values in the three phases of the study.

The raters achieved more consistency in the immediate post-training phase. In the delayed post-training phase, although the raters were still more consistent than in the pre-training phase, they had reduced consistency compared to the immediate post-training phase to a considerable extent. For a great number of the raters, the training program and feedback were pretty beneficial and brought the raters within the acceptable



	NEW 1	NEW 2	NEW 3	NEW 4	NEW 5	NEW 6	NEW 7	NEW 8	NEW 9	NEW 10	OLD 1	OLD 2	OLD 3	OLD 4	OLD 5	OLD 6	OLD 7	OLD 8	OLD 9	OLD 10
Pre-training	1.2	1.5	1.2	1.1	1.3	1.8	1.7	0.3	1.6	1	0.4	1.6	0.9	0.3	1.1	0.8	0.9	0.3	1	0.7
Immediate post-training	0.9	1	1.1	0.8	1	1.4	0.8	0.9	1.2	0.9	0.8	0.7	1.3	0.6	1.2	0.7	0.7	0.5	1.3	0.6
Delayed post-training	1.2	1.2	0.7	1.4	1.1	1.4	0.9	1.1	1	0.8	1	1.1	1.5	0.7	1.4	1	0.8	0.4	1	1.2

Fig. 10 Raters' rating consistency measures in the three phases of the study

Table 7 Effect of each facet on total score variance (delayed post-training)

No.	Facets identified in the study	Percentage effect on total score variance
1	Test taker ability	61.85
2	Rater severity	22.51
3	Group expertise	9.29
4	Test version	2.67
5	Task difficulty	3.04
6	Scale categories	0.64
		100

range of consistency after training. It was only rater OLD8 (Infit MnSq. = 0.5) who still displayed inconsistency after training. In the delayed post-training phase, although there was more consistency compared to the pre-training phase, a few more raters seem to have lost consistency compared to the immediate post-training phase. Raters OLD3 and OLD8 with the Infit Mean Square values of 1.5 and 0.4 respectively show inconsistency after training. It must be indicated that the raters who did not improve or even lost consistency after training were among the ones who were not positive about the rater training program and the feedback the raters were to be provided.

Likewise, in the previous two phases of the study, the total score variance associated with each facet was calculated to measure the effect of each facet on total score variance during the delayed post-training phase. Table 7 displays the percentage of total score variance associated with each of the facets used in the study at the immediate post-training phase. The information provided in the table shows that once again the greatest percentage of the total variance (61.85%) is attributed to the test takers' ability differences however the remaining variance (38.15%) is related to other facets including rater's severity, group expertise, test version, task difficulty, and scale categories.

In the delayed post-training phase still, a significant increase is observed towards the establishment of consistency in scoring and reduction of the influence of other intervening facets in total score variance. Here, a considerable degree of the sum of score variance is related to test takers' oral ability performance differences which shows the relative systematicity and consistency in scoring compared to the pre-training phase.

Table 8 Effectiveness of training program and feedback provision on raters' severity measures

Severity	**Successful adjustment**		**Unsuccessful adjustment**	
	N	**%**	***N***	**%**
Pre-training	4	20%	16	80%
Immediate post-training	13	65%	7	35%
Delayed post-training	10	50%	10	50%

(Pre-training × Immediate post-training) Chi-square:32.59, df=1, $p < 0.05$*

(Pre-training × Delayed post-training) Chi-square: 9.761, df=1, $p < 0.05$*

(Immediate post-training × Delayed post-training) Chi-square: 1.408, $df = 1$, $p > 0.05$

Table 9 Effectiveness of training program and feedback provision on raters' bias measures

Bias	**Successful adjustment**		**Unsuccessful adjustment**	
	N	**%**	***N***	**%**
Pre-training	13	65%	7	35%
Immediate post-training	17	85%	3	15%
Delayed post-training	15	75%	5	25%

(Pre-training × Immediate post-training) Chi-square:16.42, $df = 1$, $p < 0.05$*

(Pre-training × Delayed post-training) Chi-square:04.97, $df = 1$, $p < 0.05$*

(Immediate post-training × Delayed post-training) Chi-square:0.154, $df = 1$, $p > 0.05$

This outcome provides evidence of the ongoing efficiency of the training program in the long term. However, comparing the outcomes to the immediate post-training phase, a reduction of total score variance associated to test takers' ability and an increase of variance related to other intervening facets is observed. This outcome although still shows consistency of scoring based on test takers' oral ability, and it calls upon the gradual loss of consistency and increase of error and unsystematicity after training.

> *RQ2: To what extent was the provided feedback successful following the training program concerning severity, bias, and consistency measures?*

The following tables (Tables 5, 6, 7, and 8) demonstrate the result of training and feedback provision on *severity*, *bias*, and *consistency* measurement during the three phases for both successful and unsuccessful adjustments.

Table 8 shows the differences in the successful application of the training program and the feedback effectiveness on raters' severity reduction based on severity logit values during the three phases of the study. A pairwise comparison using a Chi-square analysis revealed that there is a considerable difference in successful severity reduction between the pre-training and the immediate post-training phase ($\chi^2_{(1)} = 32.59$, $p < 0.05$) and between the pre-training and the delayed post-training phase ($\chi^2_{(1)} = 9.761$, $p < 0.05$). However, there observed no statistically significant difference between the immediate post-training and the delayed post-training phase ($\chi^2_{(1)} = 1.408$, $p > 0.05$).

Table 9 demonstrates the same comparison but concerning biasedness. The analysis is based on the comparison of Z-score values obtained from the FACETS. The result is fairly similar to the one on severity analysis. A pairwise comparison using a chi-square analysis revealed that there is a considerable difference with respect to successful bias

Table 10 Effectiveness of training program and feedback provision on raters' consistency measures

Consistency	Successful adjustment		Unsuccessful adjustment	
	N	%	*N*	%
Pre-training	11	55%	9	45%
Immediate post-training	19	95%	1	5%
Delayed post-training	18	90%	2	10%
(Pre-training × Immediate post-training) Chi-square:23.14, *df* = 1, *p* < 0.05*				
(Pre-training × Delayed post-training) Chi-square:07.63, *df* = 1, *p* < 0.05*				
(Immediate post-training × Delayed post-training) Chi-square:0.822, *df* = 1, *p* > 0.05				

Table 11 Percentages of rater mean square fit statistics

Fit range	Pre-training		Immediate post-training		Delayed post-training	
	N	%	*N*	%	*N*	%
Fit < 0.06	4	20	1	5	1	5
0.6 ≤ fit ≤ 1.4	11	55	19	95	18	90
Fit > 1.4	5	25	0	0	1	5

reduction between the pre-training and the immediate post-training phase ($\chi^2{}_{(1)}$ = 16.42, p < 0.05) and between the pre-training and the delayed post-training phase ($\chi^2{}_{(1)}$ = 4.97, p < 0.05). However, there observed no statistically significant difference between the immediate post-training and the delayed post-training phase ($\chi^2{}_{(1)}$ = 0.154, p > 0.05).

Table 10 displays the results of consistency comparison across the three phases by comparing the data obtained from infit mean square values. The result, like what was found in the aforementioned two tables, was found. Using a chi-square analysis, there observed a significant difference in terms of successful consistency achievement between the pre-training and the immediate post-training phase ($\chi^2{}_{(1)}$ = 23.14, p < 0.05) and between the pre-training and the delayed post-training phase ($\chi^2{}_{(1)}$ = 07.63, p < 0.05). However, no statistically significant difference was obtained between the immediate post-training and the delayed post-training phase ($\chi^2{}_{(1)}$ = 0.822, p > 0.05).

As indicated before, fit statistics is used to identify which raters tended to overfit (having too much consistency) or underfit (misfit) (having too much variation) the model and at the same time to identify which raters rated consistently with the rating model. Table 11 displays the frequency and percentages of rater fit values placed within the overfit, acceptable, or underfit (misfit) categories.

Discussion

One finding of the study, which is parallel with those of (Bijani, 2010; Kim, 2011; Theobold, 2021; Weigle, 1998), also showed that not only can rater training make raters consistent in their ratings (intra-rater reliability), but also it can increase consistency among raters (interrater reliability), too. It should, however, be noted that this finding is in contrast with Davis (2019); Eckes (2008); McNamara (1996) who found that rater training can only be beneficial in promoting self-consistency but not inter-rater consistency. The reason for such discrepancy in findings might be due to the various sampling, oral tasks, or even the scoring techniques used for measuring and analyzing the data.

The findings of this study, first of all, revealed a wide variation in raters' behavior from before training to after training since they have reduced severity/leniency estimate to a high extent which made them more similar to each other. This reduction of severity estimate is more noticeable for new raters. Although severity variation among raters was reduced after training, there remained some significant severity differences among them. This, rather abnormal behavior, even after training, is due to the behaviors of some extreme raters consisting of OLD8, OLD4, OLD7 (in severity), and OLD3, OLD9, and NEW6 (in leniency) who, due to arrogance, overconfidence, or unwillingness of training program effectiveness, did not change behavior even after training and ultimately this caused overall significant variation among raters after training. In other words, those raters whose rating behavior improved very little or even got worse after the training program were those who were relatively less positive, or better to say pessimistic in their perceptions of the oral assessment rater training program. However, it is important to note that even though a causal relationship between raters' attitudes and the rating outcome cannot be formulated, it is possible to assume that if training programs are in line with the expectations and requirements of raters, they will result in more promising outcomes which will automatically result in a higher consistency with the other raters and the benchmark as well. This indicates that although training has brought raters' extreme differences within the acceptable range of severity, it could not eradicate severity variation among them. This finding is parallel with that of Stahl and Lunz (1991, cited in Weigle, 1998) who found that training cannot eliminate severity differences among raters.

Second, the training program and feedback were successful in modifying raters' fit statistics, indicating consistency among raters, after training. A considerable number of the raters who were identified as inconsistent before the training became consistent afterward. One rater (OLD8) was still identified as inconsistent after training. This might indicate that not all raters have the potential to be employed as raters and thus, according to Winke et al. (2012) and Iannone et al. (2020) should be excluded from the rating job.

The outcomes indicated that the training program was successful enough in letting the rater get closer to one another in rating and increasing their central tendency. Also, they were capable of diminishing biases compared to the pre-training phase most probably because they were provided with post-rating feedback where their biases were specifically pointed out. It also confirmed the impact of rater training on the overall consistency of raters' scoring behavior. One other possibility for the reduction of raters' biases in scoring might be on account of the fact that raters were provided with instructions that considerably provided them with explicit and clear rating procedures which probably is why little bias was observed after training. This finding is rather contradictory when compared with previous literature. That is, in terms of the reduction of raters' biases after the training program, the outcome of oral performance assessment is consistent with that of Wigglesworth (1997) who found rather the same finding regarding the reduction of bias measures after the training program. However, on the other hand, Elder et al. (2007) found a rather insignificant effect of the training program in bias reduction of raters' consequent scoring behavior.

The drastic change in rating behavior of some raters including rater OLD7 (moving from extreme leniency to extreme severity), NEW8 (moving from extreme leniency to severity), and OLD3 (moving from severity to extreme leniency) might probably be due to overgeneralization of the feedback provided. Concerning raters' fit statistics, raters who were identified as misfitting raters, according to Huang (1984, cited in Shohamy et al., 1992), could be viewed to have relative inefficiency; thus, as items on a test, to be discarded from the study. Consequently, misfitting raters had better be removed from the study; however, for the sake of examining the effectiveness of the training program, misfitting raters were kept to better observe their change of behavior in rating throughout the study. This decision has also been supported by Stahl and Lunz (1991, cited in Eckes, 2008) who stated that misfitting raters must be trained and not be excluded from the rating task.

Concerning the finding of the study in the delayed post-training phase, this study although provided promising results for the long-lasting effects of the training program, it reflected traces of gradual loss of consistency and increase of biasedness. The outcomes showed that through the lapse of time, variation gradually increases and raters tend to rate the way they rated before; however, still raters are more consistent in rating than they were before training.

Conclusions, implications, and suggestions for further research

Conclusions

One of the major findings of this study explored to what extent the training program affected the severity and internal consistency of the raters as measured by the FACETS. The outcome of data analysis through comparing pre-, and post-data demonstrated that training reduced differences in severity among raters specifically to a high extent among NEW raters, i.e., most of the raters who were identified as inconsistent before the training were no longer inconsistent afterward. The second major finding indicated that NEW raters had a broader range of severity and inconsistency than OLD ones before training. However, this was not the case after training. NEW raters tended to show less severity and higher consistency than OLD ones after training. The third finding showed that there was less variance in test takers' scores rated by the raters after training compared to the pre-training phase. Finally, the fourth finding showed that the training program helped raters realize and put the planned rating criteria into practice and helped raters modify their expectations of test-takers features and their performance ability, and their demands of the oral tasks.

The major finding was that the training program decreased yet did not eradicate the variation in severity and consistency among raters. The comparison across raters demonstrated that NEW raters had an extensive degree of inconsistency than OLD ones before the training. However, this difference was reduced after training in a way that even they became more consistent and less biased than OLD ones after training.

The outcomes of this study demonstrated that rating is still possible without training, but in order to have a reliable rating, training is essential. The primary purpose of training is to help raters articulate and justify their scoring decisions for reliable ratings. Raters, before training, differed strongly from one another concerning severity, bias, and consistency; however, following the training they diminished severity and bias to a high extent resulting in increasing the consistency in rating.

Implications of the study

Although rater training is a significant part of teacher education, it cannot make raters proficient alone. Training raters to be consistent is typically a long-lasting process since raters may not be capable of applying the techniques and strategies from training to the real scoring setting. Besides, the impacts of training might bring about changes in the delayed result. Thus, the implication is that longitudinal rater training had better be awarded before discussing the betterment of raters' scoring capability and rater variability.

The outcome of the study lead to higher degrees of interrater reliability and diminished measures of severity/leniency, biasedness, and inconsistency. However, it may turn raters identical to each other in their rating behavior. They can merely bring about higher self-consistency (intrarater consistency) among them.

Similar to the research done previously, even though rater training could assist raters to achieve higher measures of self-consistency (intra-rater reliability) and can increase interrater reliability accordingly, it cannot simply eradicate raters' differences related to their characteristics. That is, 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 through more training and individual feedback could be better paved but not thoroughly removed. The analysis outcomes of the fit statistics index of the raters demonstrated that raters are likely to increase their internal consistency in ratings through receiving training, feedback, and gaining experience.

MFRM can point out sources of raters' bias thus making assessment fairer. It can reduce the intimidation of getting either accepted or rejected based on factors that have nothing to do with their true ability. Besides, it can determine raters' bias which is the extent to which raters show interaction with either of the test versions or categories of the rating scale. The implication is that MFRM equips decision-makers with a tool to spot misfitting raters. Rating is typically an expensive and time-taking activity in which misfitting raters can invalidate test outcomes resulting in a huge loss. Therefore, although MFRM does not solve the problem, it can help provide feedback to assist raters to apply ratings more consistently.

Concerning the rather significant variation between the immediate and delayed post-training phase of the study, the outcomes of the study showed that the outcome of training might not endure long afterward. The implication is that such finding provides evidence for the requirement of ongoing training throughout the rating period letting raters regain consistency.

This study showed that raters can rate reliability, regardless of their background or level of expertise. However, rating reliability can be enhanced through training programs. The substantial rater severity/leniency differences among raters, as was also found in some previous research (e.g., Bijani & Fahim, 2011; Eckes, 2008; Theobold, 2021), have an important consequence for decision-makers that in rater training, more attention and importance shown to be dedicated to consistency within raters (intra-rater agreement) than consistency between or among raters (interrater agreement). The fact that raters reduced consistency and increased bias and severity in the delayed post-training phase, compared to the immediate post-training phase, reflects the need for assessment organizations to constantly monitor the raters based on their severity/leniency, bias, and consistency.

Suggestions for further research

This study only focused on oral performance assessment by the raters. Thus, further research could study the use of other skills (e.g., writing) and investigate raters' scoring variability including the facets used in the study on those skills as well. Besides, it did not explore the use of group oral assessment. Therefore, further studies could investigate the influence of the group oral-assessment technique on learners' performance quality and of course raters' internal agreement in scoring. On the other hand, no investigation was done regarding the differences between native and non-native speaker raters. Consequently, future studies could also investigate the differences in rating reliability as well as their behavioral variations between native speaker (NS) raters and nonnative speakers (NNS). Besides, future studies could investigate the use of raters coming from backgrounds (other than Persian language) and how they rate test takers' oral performances. Further research is required to explore the impact of the issues related to raters' and test takers' backgrounds and personalities (e.g., different first language backgrounds and language accents) on the consistency of raters in their rating.

Abbreviations

CEP	Community English Program
EFL	English as a foreign language
ETS	Educational testing system
FL	Foreign language
ILI	Iran Language Institute
MFRM	Multifaceted Rasch measurement
NNS	Non-native speaker
NS	Native speaker
SL	Second language
SLA	Second language acquisition
TOEFL	Test of English as a Foreign Language

Acknowledgements

We would like to appreciate two anonymous reviewers as well as all the raters and the students who eagerly and patiently collaborated in all the phases of this study.

Authors' contributions

Houman Bijani did the main research, collected the data and analyzed the findings. Bahareh Hashempour compiled and prepared the literature review. Salim Said Bani Orabah Wrote the discussion of the paper. Khaled Ahmad Abdel-Al Ibrahim and Tahereh Heydarnejad wrote the conclusion and revised the paper at the end. The author(s) read and approved the final manuscript.

Authors' information

Houman Bijani is an assistant professor in Applied Linguistics (TEFL) from Islamic Azad University, Science and Research Branch, Tehran, Iran. He is also a faculty member of Zanjan Islamic Azad University. He got his MA 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. Bahareh Hashempour is an M.A. holder in TEFL from Zanjan University. She has published a paper related to the application of jigsaw puzzles in language acquisition. She is a teacher and academic researcher at Safir Language Institute, Zanjan, Iran. Salim Said Bani Orabah is an Assistant Professor and the Head of English language Department, University of technology and Applied Sciences, Ibra, Sultanate of Oman. He has published some articles in valid international ISI journals. He is a Doctor of education from University of Essex in England.

Khaled Ahmed Abdel-Al Ibrahim is an associate professor of educational psychology at college of education, Prince Sattam bin Abdul-Aziz university, Saudi Arabia, and Sohag university, Egypt.

Tahereh Heydarnejad is a university lecture in at University of Gonabad, Gonabad, Iran. She published multiple papers in various international journals.

Funding

No funding was available to the present study.

Availability of data and materials

The data will be available upon requesting.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 17 June 2022 Accepted: 20 July 2022

Published online: 05 September 2022

References

Ahmadi, A. (2019). A study of raters' behavior in scoring l2 speaking performance: Using rater discussion as a training tool. *Issues in Language Teaching*, *8*(1), 195–224. https://doi.org/10.22054/ILT.2020.49511.461.

Ahmadian, M., Mehri, E., & Ghaslani, R. (2019). The effect of direct, indirect, and negotiated feedback on the tense/aspect of EFL learners in writing. *Issues in Language Teaching*, *8*(1), 1–32. https://doi.org/10.22054/ILT.2020.37680.352.

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

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.

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Peter Lang Pub Inc.

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

Davis, L. (2019). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, *36*(3), 367–396. https://doi.org/10.1177/0265532209104667.

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

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. https://doi.org/10.1177/0265532207086780.

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. https://doi.org/10.1177/0265532207071511.

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology*, *11*(1), 1–14. https://doi.org/10.3389/fpsyg.2020.0033.

Ghahderijani, B. H., Namaziandost, E., Tavakoli, M., Kumar, T., & Magizov, R. (2021). The comparative effect of group dynamic assessment (GDA) and computerized dynamic assessment (C-DA) on Iranian upper-intermediate EFL learners' speaking complexity, accuracy, and fluency (CAF). *Lang Test Asia*, *11*, 25. https://doi.org/10.1186/s40468-021-00144-3.

Hazen, H. (2020). Use of oral examinations to assess student learning in the social sciences. *Journal of Geography in Higher Education*, *44*(4), 592–607. https://doi.org/10.1080/03098265.2020.1773418.

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

Hughes, R. (2011). *Teaching and researching speaking*, (2nd ed.,). Pearson Education Limited.

Iannone, P., Czichowsky, C., & Ruf, J. (2020). The impact of high stakes oral performance assessment on students' approaches to learning: A case study. *Educational Studies*, *10*(3), 313–337. https://doi.org/10.1007/s10649-020-09937-4.

John Bernardin, H., Thomason, S., Ronald Buckley, M., & Kane, J. S. (2016). Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability. *Human Resource Management*, *55*, 321–340. https://doi.org/10.1002/hrm.21678.

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

Kwon, H., & Maeng, H. (2022). The impact of a rater training program on the TGMD-3 scoring accuracy of pre-service adapted physical education teachers. *Children*, *9*(6), 881–896. https://doi.org/10.3390/children9060881.

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

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878–902.

Liu, F., Vadivel, B., Mazaheri, F., Rezvani, E., & Namaziandost, E. (2021). Using games to promote EFL learners' willingness to communicate (WTC): Potential effects and teachers' attitude in focus. *Frontiers in Psychology*, *12,* 1-10. https://doi.org/10.3389/fpsyg.2021.762447.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54–71. https://doi.org/10.1177/026553229501200104.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331–345. https://doi.org/10.1207/s15324818ame0304_3.

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

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158–180. https://doi.org/10.1177/026553229801500202.

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

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

McQueen, J., & Congdon, P. J. (1997). *Rater severity in large-scale assessment*, *ERIC document reproduction service no. ED411303* (pp. 1–36). Center for Applied Linguistics.

Moradkhani, S., & Goodarzi, A. (2020). A case study of three EFL teachers' cognition in oral corrective feedback: Does experience make a difference? *Issues in Language Teaching*, *9*(1), 183–211. https://doi.org/10.22054/ILT.2020.51449.482.

Prieto, G., & Nieto, E. (2019). Analysis of rater severity on written expression exam using many-faceted Rasch measurement. *Psicologica*, *40*(4), 385–397.

Rezai, A., Namaziandost, E., Miri, M., & Kumar, T. (2022). Demographic biases and assessment fairness in classroom: Insights from Iranian university teachers. *Language Testing in Asia*, *12*(1), 1–20. https://doi.org/10.1186/s40468-022-00157-6.

Rosales Sánchez, C., Díaz-Cabrera, D., & Hernández-Fernaud, E. (2019). Does effectiveness in performance appraisal improve with rater training? *PLoS One*, *14*(9), 1–20. https://doi.org/10.1371/journal.pone.0222694.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, *76*(1), 27–33. https://doi.org/10.2307/329895.

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

Theobold, A. S. (2021). Oral Exams: A more meaningful assessment of students' understanding. *Journal of Statistics and Data Science Education*, *29*(2), 156–159. https://doi.org/10.1111/modl.12620.

Vadivel, B., & Beena, P. V. (2019). The impact of multimedia in English language classroom of undergraduate students in engineering colleges. *International Journal of Advanced Science and Technology*, *28*(2), 194–197.

Vadivel, B., Namaziandost, E., & Saeedian, A. (2021). Progress in English language teaching through continuous professional development—Teachers' self-awareness, perception, and feedback. *Frontiers in Education*, *6*, 757285. https://doi.org/10.3389/feduc.

Wallace, M. J. (1991). *Training foreign language teachers -A reflective approach*. Cambridge University Press.

Weigle, S. C. (1998). Using FACETS to model rater training effect. *Language Testing*, *15*(2), 263–287. https://doi.org/10.1177/026553229801500205.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14*(1), 85–106. https://doi.org/10.1177/026553229701400105.

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

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

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.