ATU PRESS

ISSUES IN LANGUAGE TEACHING (ILT)
VOL. 11, NO. 2, 255-290, December 2022
https://doi.org/10.22054/ilt.2023.63589.634 Document Type: Research Paper

Facet Variability in the Light of Rater Training in Measuring Oral Performance: A Multifaceted Rasch Analysis

Houman Bijani*
Assistant Professor of TEFL, Islamic Azad University, Zanjan Branch, Zanjan, Iran

Salim Said Bani Orabah
Ph.D. in TEFL, University of Technology and Applied Sciences, Ibra, Sultanate of Oman

Received: September 16, 2021**; Accepted:** December 28, 2022

Abstract

Due to subjectivity in oral assessment, much concentration has been put on obtaining a satisfactory measure of consistency among raters. However, obtaining consistency might not result in valid decisions. One matter that is at the core of both reliability and validity in oral performance is rater training. Recently, the Multifaceted Rasch Measurement (MFRM) has been adopted to address the problem of rater bias and inconsistency; 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 endure. Consequently, this study explored the influence of the training program and feedback by having 20 raters score the oral production, as measured by the CEP (Community English Program) test, produced by 300 test takers in three phases, i.e., before, immediately after, and long after the training program. The results indicated that training can lead to higher degrees of interrater reliability and decrease in measures of severity/leniency and biasedness. However, it did not lead the raters into total unanimity, except for making them more self-consistent. Although rater training might result in higher internal consistency among raters, it cannot eradicate individual differences. That is, experienced raters, due to their idiosyncratic characteristics, did not benefit as much as the inexperienced ones. This study also showed that the outcome of training might not endure in long terms after training; thus, it requires ongoing training, letting raters regain consistency.

Keywords: Bias, interrater consistency, intrarater consistency, multifaceted Rasch measurement (MFRM), rater training, severity/leniency

***Corresponding author's email:** houman.bijani@iau.ac.ir

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 language (SL) acquisition and foreign language (FL) learning. 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 SL and FL contexts, speaking assessment is regarded as a vital matter. Such importance calls upon valid and reliable approaches to assessing this skill (Hughes, 2011). Once the discussion about assessment and scoring is raised, attention is paid to the tools and instruments used for scoring process.

One of the most significant matters related to the scoring process is the rating scale and how it is developed and used. Many 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, Nakatsuhara & Hunter, 2020). Therefore, rater effects must be considered for suitable measuring of test takers' speaking competence. A lot of research on SL 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).

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, increases. 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 a rater group and benefit consequently, or may be scored by the severest member and experience 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. Nevertheless, the more emphasis is put on reliability, the less validity is obtained (Huang, Bailey, Sass & Shawn Chang, 2020). In other words, emphasizing higher measures of reliability do not necessarily lead to valid measurements of speaking skill. The thing that paves the way for both a reliable and valid measurement of speaking skill is rater training.

The Multifaceted Rasch model introduced by Linacre (1989), 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 (Lumley & McNamara, 1995; Ahmadian, Mehri, & Ghaslani, 2019). Pioneers of the Rasch technique in assessment argue that it is impossible to train raters to obtain the same degree of severity (Lunz, Wright, & Linacre, 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, as Lumley and McNamara (1995) state, rater variation could be found out with respect to severity and random error; thus, training and even retraining are suggested for those raters who are spotted as misfitting by the Rasch technique (Lunz, Wright, & Linacre, 1990) in order 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 own subjective judgments in short intervals and as a result, alter their rating approaches in the long run accordingly.

Statement of the Problem and Purpose of the Study

Nevertheless, much of the research done up to now has explored the use of FACETS on just a couple of facets. For instance, research has been done on rater's severity or leniency on test takers (Lynch & McNamara, 1998), task types (Wigglesworth, 1997), and specific rating time (Lumley & McNamara, 1995). However, no research has incorporated the facets of test takers' ability, including the 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), rater variability can still be 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). In addition to that, little research has explored the duration of rater training effects (Bijani, 2010). There are

researches 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.

Therefore, this study aimed to focus on the feedback that the raters received during training in relation to its impact on their 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. This study also intended to analyze each rater's rating behavior so that it would provide feedback to the raters 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:

RQ1: How much of test takers' total score variance can be accounted by each facet, i.e., test takers' ability, rater severity, raters' group expertise, task difficulty, test version, and rating scale?

RQ2: To what extent was the provided feedback effective following the training program regarding reducing severity, bias, and increasing consistency measures?

LITERATURE REVIEW

In scoring SL speaking performance, rater variability has been identified as a potential source of measurement error, which might interfere with the measurement of test-takers' true speaking ability (McNamara 1996; Reed & Cohen, 2001). Therefore, rater effect is required to be taken into

consideration in order to measure test takers' speaking ability appropriately. Many studies on second language speaking assessment by raters have focused on sources of speaking variability. These variables include rater severity, interaction with other aspects of the rating situation, and internal self-consistency (Lumley & McNamara, 1995; Lynch & McNamara, 1998; Wigglesworth, 1997).

Other studies on rater variability have investigated raters' decision-making processes qualitatively, often by means of verbal reports (Brown, 2005). According to Luoma (2004) raters' verbal report data are analyzed to examine which features of test takers' responses and of course the scoring criteria, raters pay attention while scoring speaking performances. Generally, human judgment is used to assign scores in oral assessment. However, in case critical decisions are made on the basis of such ratings, it is essential to ensure the accuracy and fairness of the assigned scores. As a result, rater selection, training and monitoring procedures should be chosen for the intention of minimizing the effect of rater inaccuracy and bias. In the absence of rater agreement, raters do not tend to award equal scores to the same performances; thus, severity, which is the possibility of awarding lower scores by raters, and leniency, which is the reverse aspect, increases (Davis, 2016). This will turn assessment into a lottery and yield unreliable and invalid results because different raters might score a particular test taker differently. That is, a test taker may be privileged by being assessed by the most lenient member of the rating group or be disadvantaged by being scored by the severest rater in the group.

One very useful way to address the above problem is double rating in which the scores assigned by at least two raters are averaged. The process of estimating the reliability of such scoring is referred to as inter-rater reliability or internal consistency. Internal consistency, which is the target of rater training, is also closely related to the use of a particular rating scale (Davis, 2016). In other words, internal self-inconsistency may happen because raters do not have complete understanding of a given rating scale. Since self-consistency, according to Lumley and McNamara (1995), often

cannot be obtained by rater training, it is assumed that what is important in obtaining consistency is how well a rater masters the guidelines of a special rating scale. Thus, the focus of rater training should be determined in a wider scope beyond a single source of rater variability (i.e., internal consistency) to the raters' appropriate use of scoring criteria as defined in the rubric (Lumley & McNamara, 1995). In this respect, the reliability, validity and practicality of any oral test are clearly of chief concern, and thus related to all aspects of the issues discussed above. The training of raters is another area of reliability, which will be discussed in the later parts of this study.

Another area of concern is test method facets, which may affect the reliability and validity of research studies because of their impacts on scores (Theobold, 2021). Test method facets include not only tasks, but also raters and their degree of training in the use of the rating scale plus the testing format and the rating criteria. In recent years, direct oral examinations have become a standard practice in assessing the speaking skills in both first and second languages. Since such tests require subjective evaluations of speaking quality, a great deal of emphasis in research studies has been placed on achieving an acceptable level of inter-rater reliability in order to show that spoken language can be scored as fairly and consistently as possible. However, this emphasis on reliability has been at the expense of decreasing test validity (Weigle, 1998); that is, the procedure for achieving higher reliability may not necessarily lead to valid judgments of speaking quality. Consequently, the issue at the heart of both reliability and validity in performance assessment (e.g., speaking assessment) turns out to be 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. A number of scholars including (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, Czichowsky & Ruf, 2020). A fairly substantial amount of literature, commencing with the research done by Frawley and Lantolf (1985) 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 language SL 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.

In performance assessment, rater training has also been referred to as well, 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 in fulfilling this objective. For example, even if we train raters to assign precisely similar scores to test takers, which is obviously farfetched, there still remains concerns regarding the interpretability of such scores. Linacre further argues and addresses the solution to the problem of the interpretability of the score through a rather recent approach to analyzing raters' scores known as the Multifaceted Rasch measurement. This way, as he reiterates, both the reliability and the validity of the given scores are established.

METHOD

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 learning English at the Iran Language Institute (ILI).

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 and Master 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, in order to keep the data provided by the raters confidential, their names and identities were anonymized through 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 including (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 1.

Table 1: 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 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.

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

A. Raters with no or fewer than two years of experience, outlined by McNamara and Lumley (1997), 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.

Instrumentation

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 listenings. For tasks 2 (*Narration Task*) and 5 (*Exposition Task*) the test takers are needed to give 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 in a more valid and reliable way. 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 on the basis of *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

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 among the three groups. Meanwhile, the raters were awarded the background questionnaire prior to 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 researchers commenced the speaking test. It is worthy to mention 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-/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 the 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 6 band analytic rating scale, to the researchers.

Rater Training Procedure

Once the pre-training phase was over, the raters took part in a training or norming session during they got familiar with the oral tasks and the rating scales. They also had the opportunity to practice the instructed materials provided with a number of sample responses. The researchers 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 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 feedback was based on the raters' use of the rating scale, their evaluation of each descriptor of the scale and their possible severity/leniency and bias each could have during their judgment. 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. Due to the fact that each rater possesses a different rating ability and rating behavior, it was essential that each rater be provided with feedback individually.

Immediate Post-training Phase

Immediately following the rater training program discussed above, when the raters got the required skill in rating speaking ability, the tasks of both versions of the test 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 to collect data. It is again stressed that the oral tasks were assessed using the ETS rating scale.

Delayed Post-Training Phase

Exactly two months (as suggested by McNamara, 1996) after the immediate post training data collection, the fifth phase of data collection procedure was done. In this phase, the last third of the test takers (including 100 students) were tested from whom to obtain data. The raters were provided with the collected data to rate on the basis of the knowledge they had already gained during the rater training program two months before. As it was mentioned above, the last third of the test takers, including 100 students, took part to provide oral performances which were rated by the raters. The aim was to observe the delayed impact of the training program on raters. The expectation was that the raters were still consistent in their rating. The results of the analyzed data for this step of the study would show the delayed effectiveness of the training program on raters and also the degree of inter-rater reliability.

Data Analysis

In order to address the research questions indicated in this article, the researchers of the study used a pre-post, quantitative research design

investigating the raters' development over time with respect to scoring second language speaking performance (Cohen, Manion & Morrison, 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).

Quantitative data (i.e., raters' scores on the basis of an analytic rating scale) were gathered and analyzed with MFRM during three scoring sessions. 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

The data at the pre-training phase of the study were analyzed, and the FACETS variable map representing all the facets was obtained. In the FACETS variable map, presented in Figure 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 which 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.

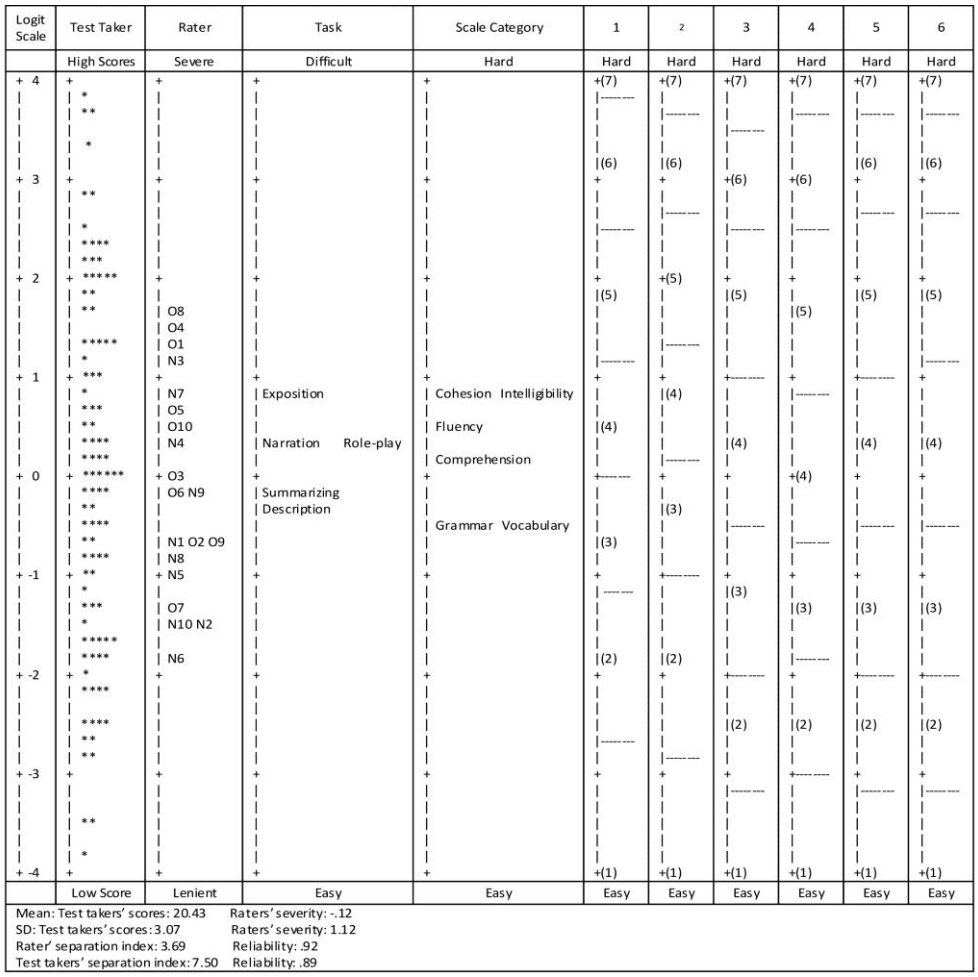


Figure 1: FACETS variable map (pre-training)

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 at the pre-training phase of this research.

The *third column (Rater)* displays raters with regard to their severity or leniency estimates in scoring test takers' oral proficiency. Since there were 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. Obviously, 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** 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 six to eleven (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 which 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 to +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 each facet, i.e., test takers' ability, rater severity, raters' group expertise, task difficulty, test version, and rating scale?

A FACETS program enables us to determine how much each score variance is attributed to the facets employed. Accordingly, one more data analysis was done in order to measure to what extent the total score variance is associated to each of the facets identified in this study. Table 2 shows the percentage of total score variance associated to 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 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.

Table 2: 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

The rather high percentage of total score variance, other than that of test takers' capability at the pre-training phase calls up on the caution to be taken with regard to 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 for a significant extent of total test variance (26.13) which indicates that there is likelihood towards 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 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.

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

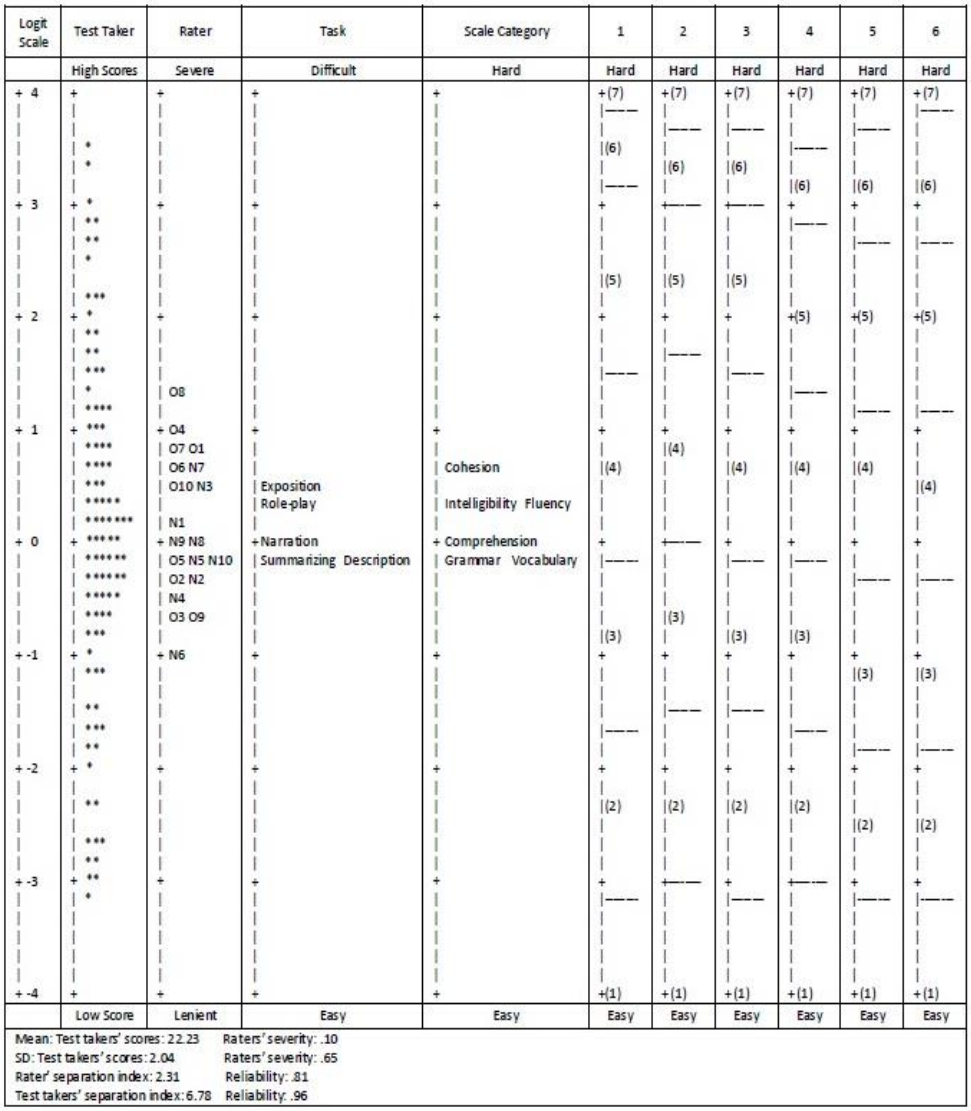


Figure 2: FACETS variable map (immediate post-training)

The *second column (Test Taker)* displays the 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 with respect to raters' scoring of their oral proficiency level.

The *third column (Rater)* displays raters with regard to their 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. Likewise, 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 with regard to severity/leniency indices. In other words, they rated more similarly with regard to 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** measured to be the most severely category (logit value = 0.58) for raters to use whereas **Grammar** was the least severely 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 3 displays the percentage of total score variance associated to 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.

Table 3: 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

The considerable increase in 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 at the immediate post-training phase. The scoring procedure moved towards establishment of consistency in scoring in a way that a majority of score variance was associated to test takers' performance ability differences.

The data at the delayed post-training phase of this research were analyzed, and the FACETS variable map representing all the facets was obtained. The FACETS variable map, displayed in Figure 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 takers proficiency ranges from 3.70 to -3.53 logits, with a logit distribution of 7.23.

The *third column (Rater)* displays raters with regard to 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 the previous two phases of the study, raters, as shown in the figure, seem 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 at 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 of severity logit measure compared to the immediate post-training phase (2.31 logits) reflects the raters' tendency in moving gradually to their own 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.

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** 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).

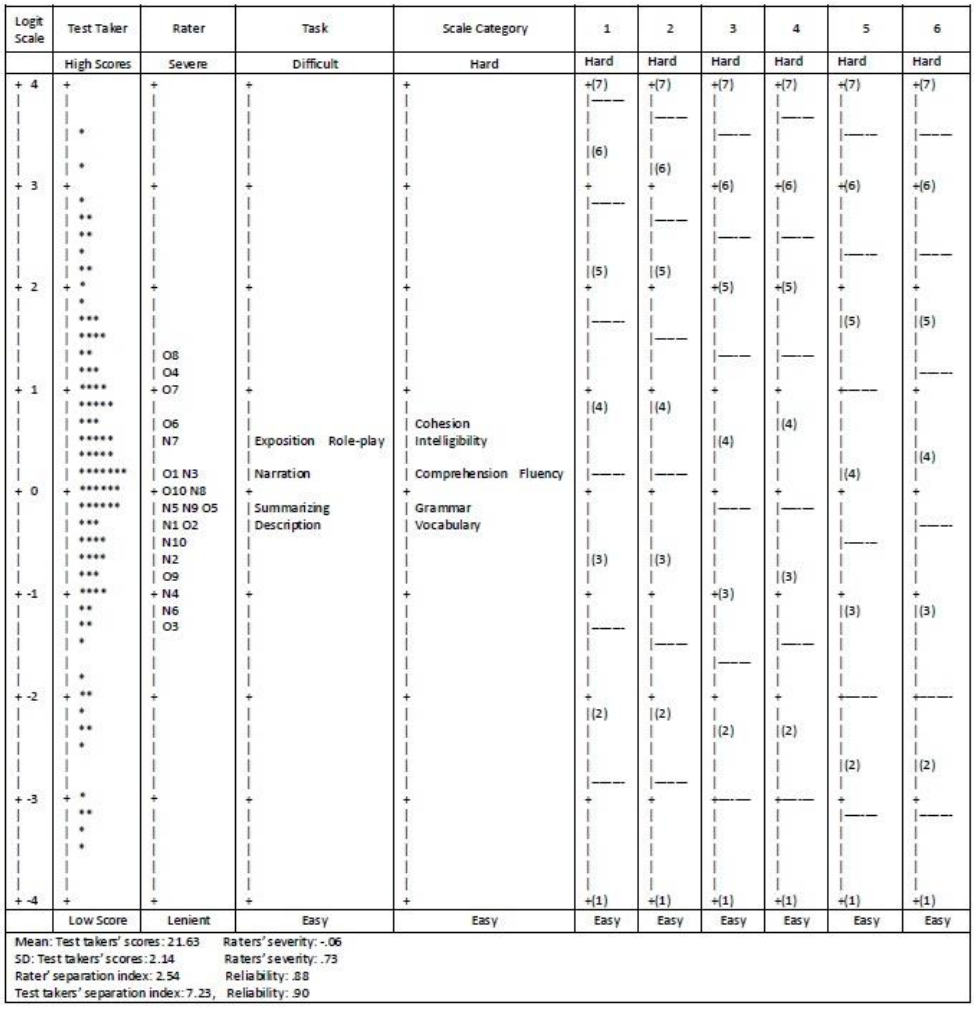


Figure 3: FACETS Variable Map (Delayed post-training)

Figures 4 to 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 it 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.

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.

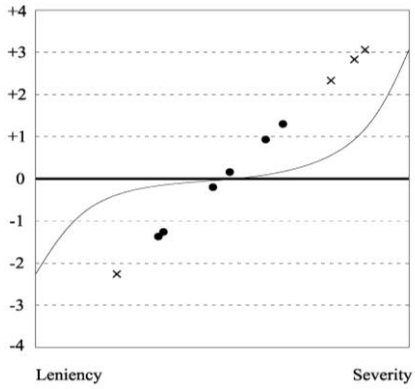


Figure 4: OLD raters' bias interaction (Pre-training)

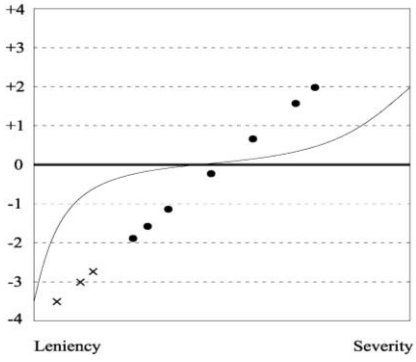


Figure 5: NEW raters' bias interaction (Pre-training)

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

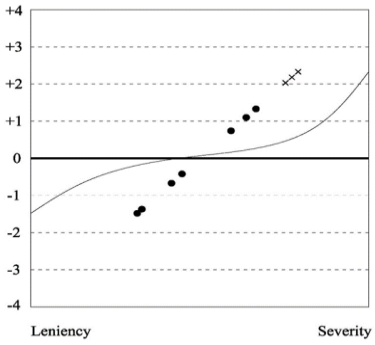


Figure 6: OLD raters' bias interaction (Immediate post-training)

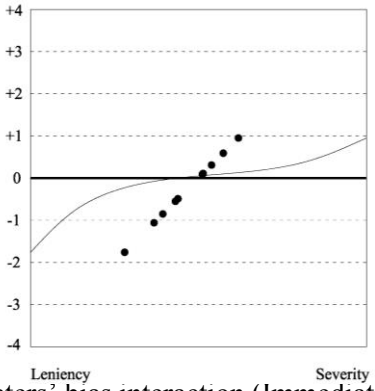


Figure 7: NEW raters' bias interaction (Immediate post-training)

Delayed post-training: there were 1 significant bias for NEW raters which 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 biases among which 3 were identified as significantly severe and 1 lenient. One rater was on the borderline of severity measure.

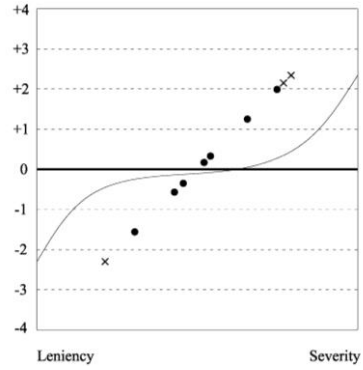


Figure 8: OLD raters' bias interaction (Delayed post-training)

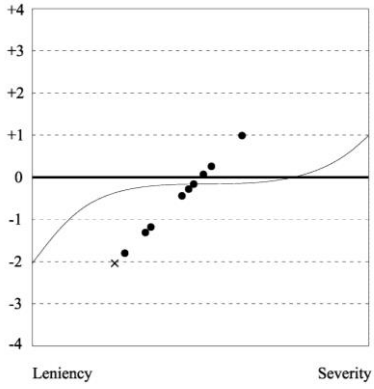
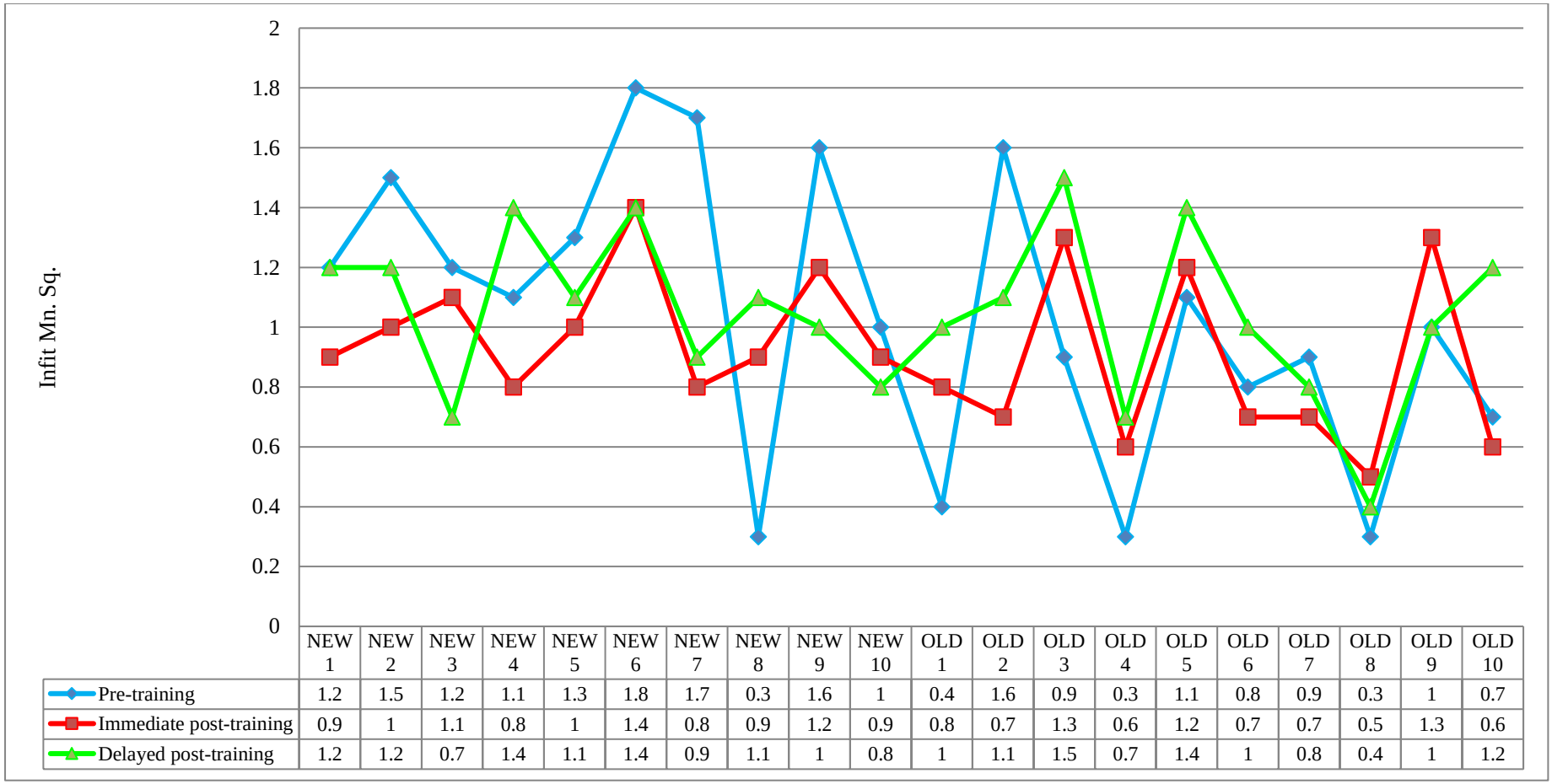


Figure 9: NEW raters' bias interaction (Delayed post-training)

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 as the acceptable range (Wright & Linacre, 1994). The following figure (Figure 10) plots graphically the change of raters' consistency in rating using infit mean square values in the three phases of the study.



	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

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

It is clear that 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 the pre-training phase, they experienced a reduction in their 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 range of consistency after training. It was only rater OLD8 (Infit MnSq. = 0.5) who still displayed inconsistency after training. At the delayed post-training phase, although there were 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 having Infit Mean Square value of 1.5 and 0.4 respectively showing 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, the previous two phases of the study the total score variance associated to each facets was calculated to measure the effect of each facet on total score variance at the delayed post-training phase. Table 4 displays the percentage of total score variance associated to 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.

Table 4: 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

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

RQ2: To what extent was the provided feedback effective following the training program regarding reducing severity, bias and increasing consistency measures?

The following tables (Tables 5 to 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 5 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 with regard to successful severity reduction between the pre-training and the immediate post training phase ($X^2_{(1)} = 32.59$, $p<0.05$) and between the pre-training and the delayed post-training phase ($X^2_{(1)} = 9.761$, $p<0.05$). However, there observed no statistically significant difference between the immediate post-training and the delayed post-training phase ($X^2_{(1)} = 1.408$, $p>0.05$).

Table 5: 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 6 demonstrates the same comparison but with respect to 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 reduction between the pre-training and the immediate post training phase ($X^2_{(1)} = 16.42$, $p<0.05$) and between the pre-training and the delayed post-training phase ($X^2_{(1)} = 4.97$, $p<0.05$). However, there observed no statistically significant difference between the immediate post-training and the delayed post-training phase ($X^2_{(1)} = 0.154$, $p>0.05$).

Table 6: 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, $p>0.05$				

Table 7 displays the results of consistency comparison across the three phases of through 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 ($X^2_{(1)} = 23.14$, $p<0.05$) and between the pre-

training and the delayed post-training phase ($X^2_{(1)} = 07.63$, $p<0.05$). However, no statistically significant difference was obtained between the immediate post-training and the delayed post-training phase ($X^2_{(1)} = 0.822$, $p>0.05$).

Table 7: 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, $p>0.05$				

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 8 displays the frequency and percentages of rater fit values placed within the overfit, acceptable, or underfit (misfit) categories.

Table 8: 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

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 own ratings (intra-rater reliability), but also it can increase consistency among raters (interrater reliability). 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 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 still 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 to the oral assessment

rater training program. However, it is important to note that in spite of the fact that the causal relationship between raters' attitudes and the rating outcomes 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 will 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 totally eradicate severity variation among them. This finding is parallel with that of Stahl and Lunz (1991, cited in Weigle, 1998) and Mohd Noh and Mohd Matore (2022) who found that training cannot eliminate severity differences among raters.

Second, the training program and feedback were successful in modifying the raters' fit statistics, indicating consistency among raters after the training. A considerable number of the raters who were identified inconsistent prior to the training became consistent afterwards. One rater (OLD8) was still identified as inconsistent after training. This might indicate that not all raters have the potentiality to be employed as raters and thus, according to Winke, Gass & Myford (2012) and Iannone, Czichowsky and Ruf (2020) should be excluded from the rating job. The findings displayed that a little change in raters' degree of severity can have a great effect on test takers' relative position with regard to their oral performance ability. Therefore, students must be constantly monitored and checked for their true ability.

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 due to the fact that they 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 about the reduction of raters' biases in scoring might be on account of the fact that raters were provided with instructions which considerably provided them with explicit and clear rating procedure 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 Barkhuizen, Knoch and Randow (2007) found rather insignificant effect of the training program in bias reduction of raters' consequent scoring behavior.

The drastic change of the rating behavior of a number of 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. With respect to raters' fit statistics, raters who were identified as misfitting raters, according to Huang (1984, cited in Shohamy, Gordon & Kraemer, 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, 2005) who stated that misfitting raters must be trained and not be excluded from the rating task.

With respect to the finding of the study at the delayed post-training phase, this study although provided promising results for the long-lasting effectiveness 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. This outcome is consistent with that of Hazen (2020) who found higher measures of consistency still in long-terms following rater training programs.

One of the major findings of this study pertained to the extent to which 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-, 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 inconsistent prior to the training were no longer inconsistent afterwards. The second major finding indicated that NEW raters had a broader range of severity and inconsistency than OLD ones prior to training. However, this was not the case after training. 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 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. These results confirmed those of previous studies (e.g., Bijani, 2010; Hazen 2020; Theobold, 2021; Weigle, 1998)

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 prior to 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.

CONCLUSION

The outcomes of this study demonstrated that rating is still possible without training, but in order to have 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 with respect to severity, bias and consistency; however, following training they diminished severity and bias to a high extent resulting in an increase in consistency in rating.

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, longitudinal rater training had better be awarded prior to 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 exactly 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 (e.g., Bijani, 2010; Hazen 2020; Theobold, 2021; Weigle, 1998) 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' individual 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.

With respect to 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 afterwards. This finding provides evidence for the requirement of ongoing training throughout the rating period letting raters regain consistency.

This study showed that raters are able to 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 at 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.

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 to 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 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' background and personality (e.g., different first language background and language accents) on the consistency of raters in their rating.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Houman Bijani http://orcid.org/0000-0002-4305-7977
Salim Said Bani Orabah http://orcid.org/0000-0003-4792-9346

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: 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.* Frankfurt: Peter Lang Pub Inc.

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

Davis, L. (2019). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 36*(3), 367-396. https://doi.org/10.22055/RALS.2020.15418

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies*. Oxford: 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. and 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.2020.00330

Frawley, W., & Lantolf, J. P. (1985). Second language discourse: A Vygotskyan perspective, *Applied Linguistics, 6*(1), 19-44. https://doi.org/10.1093/applin/6.1.19

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.). London: 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

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

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

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878-896.

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. 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.* London: Longman.

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

McQueen, J., & Congdon, P. J. (1997). *Rater severity in large-scale assessment*. Washington, DC: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED411303, pp. 1-36).

Mohd Noh, M. F., & Mohd Matore, M. E. E. (2022). Rater severity differences in English language as a second language speaking assessment based on rating experience, training experience, and teaching experience through many-faceted Rasch measurement analysis. *Frontiers in Psychology, 13*(4), 94-108. https://doi.org/10.3389/fpsyg.2022.941084

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.

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/10.1080/26939169.2021.1914527

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