Bijani, *Cogent Education* (2018), 5: 1460901
https://doi.org/10.1080/2331186X.2018.1460901

cogent education

TEACHER EDUCATION & DEVELOPMENT - ENGLISH AS A FOREIGN LANGUAGE | RESEARCH ARTICLE

Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training

Houman Bijani[1]*

Received: 16 August 2017
Accepted: 29 March 2018
First Published: 04 April 2018

*Corresponding author: Houman Bijani, Department of English Language Teaching, Islamic Azad University, Zanjan, Iran
E-mail: houman.bijani@gmail.com

Reviewing editor:
Jennifer Mitton Kukner, Saint Francis Xavier University, Canada

Additional information is available at the end of the article

Abstract: Rater variability has always been identified as an important source of measurement error in performance assessment, especially for oral proficiency tests. Rater training is commonly used as a means for compensating various sources of rater variability and adjusting their assessment quality. However, there is little research regarding the nature of training programs and raters' perception using both a qualitative and a quantitative research design. Despite previous data on test takers' reactions to oral test performance, there is little research regarding the application of test feedback and raters' perceptions of the given feedback on their scoring performance and its probable usefulness on raters' ratings. Twenty raters rated 300 test takers' oral performances before and after a training program and their perceptions, attitudes, expectation, and evaluations were identified via questionnaires, interviews, and observations. The findings of qualitative and quantitative data analyses demonstrated that training programs are quite useful in satisfying their attitudes, perceptions, and evaluations about it. This will definitely result in the reduction of their severity and biases and increase in their consistency level. Besides, informing raters of the goals of performance assessment in training programs will

Houman Bijani

ABOUT THE AUTHORS

Houman Bijani is a PhD holder in TEFL from Azad University, Science and Research Branch, Tehran, Iran. He is also a faculty member of Zanjan Azad University. He got his MA in TEFL from Allameh Tabatabai University as a top student. He has published several research papers in scholarly national and international language teaching and assessment journals. His areas of interest include quantitative assessment, teacher education, and language research. The current research paper is a part of a wider study entitled oral performance assessment: the use of FACETS in raters' rating biasedness which was conducted for the fulfillemet of the requirement of PhD study.

PUBLIC INTEREST STATEMENT

Differences among raters cause variability in speaking assessment. This variability could be paved by rater training programs. However, there is little research dealing with the perceptions and attitudes of raters about rater training programs and whether their attitudes and perceptions could have any impact in their rating performances. In this research, the perceptions and attitudes of 20 raters about the rater training program was obtained through questionnaires and interviews. The data were then analyzed qualitatively and quantitatively indicating the usefulness of the impact of training programs in the reduction of raters' consistency in scoring. The findings also indicated that the more positive the raters feel toward the training program, the better they can use the instructions in their subsequent ratings. Such findings indicate that decision-makers should not be concerned about raters' expertise levels, but they should establish better rater training programs to increase rater consistency in their assessments.

result in less halo effect. Finally, those having positive attitudes toward rating feedback were able to incorporate it more successfully in their rating and thus achieved more consistency and less biasedness in their subsequent ratings. Consequently, decision-makers should not be concerned about raters' expertise levels, but they should establish rater training programs to increase rater consistency and reduce their biases in measurement.

Subjects: Educational Research; Education Studies; Research Methods in Education; Continuing Professional Development; Language Teaching & Learning

Keywords: bias; halo effect; oral assessment; rater's perception; rater training

1. Introduction

The ability to speak in a second language is widely recognized as an important skill for educational, business, and personal reasons (Kim, 2015). Therefore, with the significance attached to speaking in second and foreign language contexts, testing speaking is considered an important issue. According to Fulcher (2003) second-language speaking tests "have always been thought of as important, but they were too unreliable and impractical to use in the kind of large-scale testing that emerged in the 1920s with a rapidly expanding educational system" (p. 16). When discussing scoring procedures, we are concerned with the way in which the scoring system, most commonly the rating scale, is developed and used. Mostly and in many oral performance tests, students' performances are rated subjectively via employing a rating scale by the use of the descriptors based on which a rater assigns a scoring number (Kuiken & Vedder, 2014). Because such tests require subjective evaluations of speaking quality, a great deal of research emphasis 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 constantly as possible. However, this emphasis on reliability has been at the expense of decreasing test validity (Chalhoub-Deville, 1995); that is, the procedure for achieving higher reliability may not lead to valid judgments of speaking quality. Therefore, one issue which is at the heart of both reliability and validity in essay scoring is that of rater training (Fulcher, 2003). However, it is clear that training does not have the only, or even the most significant, effect on the rating of speaking ability of test takers. A number of possible sources of rater disagreement have been studied and explored in the literature of speaking assessment (e.g. Brown, 2003; In'nami & Koizumi, 2016) which can have serious impact on raters' assessment both positively and negatively.

With regard to the issue of the validity and oral test validation, raters' perceptions are one of the outmost important factors. Winke, Gass, and Myford (2012) identifies factors such as introversion/extroversion, intelligence, experience, education, social norms, and willingness to communicate as factors that may influence raters' affective reactions to the oral testing situation and consequently their scorings. The underlying problems dealing with the nature of communication in oral testing is the choice of topic (Ling, Mollaun, & Xi, 2014) which is also related to affective aspects of testing. These factors may systematically influence the behavior of raters under test conditions independently of the oral variable traits that the researcher aims to measure. It is therefore the responsibility of the test researcher to ensure what factors account for prior to the interpretation of results. Kenyon and Stansfield (1991) suggested field testing of tasks along with the use of questionnaires to elicit raters' perceptions to determine good and poor tasks. In an attempt, they scaled a number of oral speaking tasks, used in the ACTFL oral assessment, based on their functions. Then, by the use of a Rasch partial credit model they assessed the difficulty of a number of tasks. Although they found a reasonable correlation between the suggested difficulty level and the assessment of difficulty by raters, this is far from testing tasks on students and assessing task difficulty from scores.

2. Literature review

2.1. Rater variability

In scoring second-language speaking performance, rater variability has been identified as a potential source of measurement error which might interfere with the measurement of test takers' true speaking ability (In'nami & Koizumi, 2016; Reed & Cohen, 2001). Therefore, rater effects are required to be taken into consideration in order to measure test takers' speaking ability appropriately. The reliability of a rating scale is also dependent on the raters who operate it (Overall & Magee, 1992). So, raters, as an additional source of measurement error, can be a magnificent variable on test scores. Rater variability has been shown to demonstrate itself in many different ways. The main differences, as Brown (2003) asserts, could be variation in their interpretations of the rating scale, level of severity, impressions toward test takers (halo effect), test takers' backgrounds such as their gender and background knowledge. Most importantly, it is now well proved that raters vary with respect to their severity in their judgments of test takers' performance ability (McNamara, 1995). McNamara and Adams (1991) in a study found differences in raters' behaviors depending on various groups of test takers on different tasks in use. This is something that was referred to as *bias* by Linacre (1989).

2.2. Rater background

One important, related rater feature that has been demonstrated to influence test takers' test scores is rater background. Various groups of raters may differ in the judgment of learners' second-language ability depending on their background and the criteria they apply (Barrett, 2001). Several studies have found differences between inexperienced and experienced raters in their scorings and the use of rating strategies (Attali, 2016; Bijani & Fahim, 2011; Davis, 2016; Khabbazbashi, 2017; Knoch, Read, & von Randow, 2007; Kyle, Crossley, & McNamara, 2016; Winke et al., 2012). Davis (2016) compared the ratings of trained and untrained raters from two various backgrounds: experienced English teachers and non-teachers. They found that training was a more significant variable than background in terms of reliability. However, they did not report any differences with regard to the overall differences between groups in rater severity. In a similar study, Bijani (2010) and Khabbazbashi (2017) found that experienced teachers were severer than novice ones. However, Bijani (2010) found training programs influential in the reduction of severity measures. The importance is to investigate the impact of rater expertise in their perceptions of training programs.

2.3. Rater expectation

Another important rater variable is that of raters' expectations about students' speaking. Kim (2015) stated that raters' expectations may be as important as the quality of the performance itself in the score assigned. Some scholars found that raters gave higher scores to the same speaking performances when they were told that they were uttered by honor students (Knoch, 2011; Nakatsuhara, 2011; Van Moere, 2012). However, Van Moere (2012) notes that such results are not surprising, given that raters' expectations and experience are important parts of the rating process. Rater training is commonly used as a means for compensating various rater backgrounds and adjusting rater expectations to reduce the variability in this respect. Similarly, Nakatsuhara (2011) found overall modification in raters' expectations as a result of training who found that training modifies raters' expectations and understandings of rating. Yet, Knoch (2011) in her study found mismatches regarding raters' perceptions and the impact of training on their scoring behavior.

However, although training has been discussed in the literature of speaking assessment (Attali, 2016; Davis, 2016; Hamilton, Reddel, & Spratt, 2001; Kuiken & Vedder, 2014), there is little research regarding the nature of training programs and raters' perceptions of the training program using both a qualitative and a quantitative research approach. Besides, few studies, if any, have investigated the differences between experienced and inexperienced raters in their perceptions of training programs and the effect of which on them. Despite previous data on test takers' reactions to oral test performance, there is little research regarding the application of test feedback and of course the raters' perceptions of the given feedback on their scoring performance and its probable usefulness on raters' subsequent ratings. Besides, very few studies, if any, have used a mixed-method approach,

known as the third methodological movement (Tashakkori & Teddlie, 2003), having a mixture of qualitative and quantitative research methods and techniques in a single-study analyzing raters' attitudes and perceptions. Therefore, this study aims to analyze a variety of rater background characteristics including their background rating variables (i.e. raters' previous rating experiences) that raters bring into the rating context that affects scoring test takers' speaking performance along with the impact of their attitudes and perceptions of oral assessment and rater training programs on the validity of their evaluations. Consequently, the following research question can be formed:

RQ1: What are the raters' perceptions, attitudes and evaluations of the oral assessment test, test takers' performance and the training program both before and after training? *(Qualitative)*

RQ2: Is there any significant difference between their perceptions before and after training? *(Quantitative)*

3. Methodology

3.1. Participants

Three hundred adult Iranian students of English as a Foreign Language (EFL), including 150 males and 150 females, ranging in age from 17 to 44 participated in the study voluntarily as test takers. The students were selected from intermediate, upper-intermediate, and advanced levels studying at the Iran Language Institute (ILI).

Twenty Iranian EFL teachers, including 10 males and 10 females, ranging in age from 24 to 58 participated in this study as raters. These raters were undergraduate and graduate in English language related fields of study, teaching in different universities and language institutes. The raters in this study were selected based on availability at the time of the study and purposeful sampling (Dörnyei, 2007); that is, those who have already got the qualifications and of course were willing to participate took part in this study. It should also be stated that all the raters had high levels of English language proficiency although none was a native speaker of English language. In order to fulfill the requirements of this study, the raters were classified into two groups of experienced raters and inexperienced ones to investigate the similarities and differences among them and the likelihood advantages of one group over the other one. Moreover, each rater was referred to by assigning a number from 1 to 10 to promise their anonymity and preserve their confidentiality of the data reported by them. A background questionnaire, adapted from McNamara and Lumley (1997), eliciting the following information including (1) *demographic information*, (2) *rating experience*, (3) *teaching experience*, (4) *rater training*, and (5) *relevant courses passed* was given to the raters. Therefore, the raters were divided into two levels of expertise on the basis of their experiences outlined below.

(1) Raters who had no or less than two years of experience in rating and receiving rater training, and had no or less than five years of experience in teaching and passed less than the four core courses related to ELT major (pedagogical English grammar, phonetics and phonology, second-language acquisition, and second-language assessment). Hereinafter, we call these raters as NEW.

(2) Experienced raters who had over two years of experience in rating and receiving rater training, and over five years of experience in teaching and passed all the four core courses plus at least two selective courses related to ELT major. Hereinafter, we call these raters as OLD.

A more important reason for choosing these groups of expertise is to investigate any differences between experienced and inexperienced raters in terms of their attitudes and perceptions of the training program, how they approach the task of oral assessment and how they are affected by the rating process when assessing test takers. It is noteworthy to indicate that in order to eliminate 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. It is noteworthy to indicate that in

Table 1. 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

order to meet ethical considerations, the test takers and the raters had entitled the researcher the permission for the publication of any research document out of this study. Table 1 displays the summary characteristics of the raters participating in the study.

3.2. Instruments

3.2.1. Raters' pre-training questionnaire

The pre-training questionnaire which was used in this study aimed to focus on individual rater's degree of attitude, perceptions, feeling, reaction, and expectations from the training program. This questionnaire had originally been developed by Elder, Barkhuizen, Knoch, and von Randow (2007) to assess the degree of raters' attitude toward the training program. The original questionnaire consisted of five items; however, to make it more suitable for this study, it was modified and some more items were added to fit the requirements of this study. The reliability of the adapted questionnaire was measured through statistical data analysis and several pilot studies ($r = 0.81$); however the details were not mentioned for the sake of keeping the brevity of the study. Thus, the final version of the rater's pre-training questionnaire consisted of 13 items.

3.2.2. Raters' pre-training interview

The raters participating in this study were also interviewed after having filled out the pre-training questionnaire. The interview is designed to provide additional, more extensive feedback about their attitudes and feelings, tendency to accept authorities' comments, their expectations of the training program, and suggestions for the development of the assessment. The addition to the raters' pre-training interview was to fulfill the requirements of a triangulated qualitative research in order to increase the validity of the study. The interview question items were developed by the researcher to ensure the validity of the data collected by questionnaire.

3.2.3. Raters' post-training questionnaire

The post-training questionnaire which was used in this study was aimed to focus on individual's attitudes, perceptions, effectiveness and evaluation, efficacy, the provided feedback, and further developments of the training program. This questionnaire was also originally developed by Elder et al. (2007) to assess the degree of raters' attitudes, evaluation and the effectiveness of the training program. The original questionnaire consisted of six items; however, to make it more suitable for the present study, it was modified and thus it consisted of 15 items. Similarly, the reliability of the adapted questionnaire was measured through statistical data analysis and several pilot studies ($r = 0.77$).

3.3. Procedure

3.3.1. Pre-training phase

Prior to collecting any data from the test takers, the raters' background questionnaire was given to the raters to fill out before starting to run the test tasks and collect data. The aim of having the raters fill out the raters' background questionnaire sheets was to enable the researcher to classify them into the two groups of rating expertise i.e. inexperienced raters and experienced ones. It is reiterated that individual tasks are assessed using appropriate scoring criteria including fluency, grammar, intelligibility, vocabulary, cohesion, and comprehension consisting each criterion of a set of seven descriptors ranging from 1 to 7 using the ETS (2014) scoring rubric.

All the raters participating in this study were given one week to submit their scorings, Moreover, the videotaped recordings of the oral assessment settings were awarded to the raters to assist them observe aspects of oral performance including metalinguistic behaviors and body language. Having done so, they were provided with the pre-training questionnaire to fill out and indicate their attitude, feelings, opinions about the comparability of the two test versions, and expectations of the training program. Meanwhile, since some raters might not answer all the items of the pre-training questionnaire in details, they were also interviewed to get their ideas regarding the training program prior to running it. It must be reiterated that all the interviews were audio-taped and qualitatively analyzed along with the raters' pre-training questionnaires.

3.3.2. Rater training procedure

After the pre-training scoring stage, the raters participated in a training (norming) session in which the speaking tasks and the rating scale were introduced and time was given to practice the instructed material with some sample responses. The raters were provided with information specifically related to scoring procedure due to the fact that the aim of the training program was to familiarize all raters of both levels of expertise with salient features of rating when dealing with students' second-language speech samples. An attempt was made to minimize the contrasting attitudes and biases the raters might have brought with them to their assessment via providing them with assessment criteria and performance samples as well as requiring the raters to take part in the training program. Previously recorded responses were played and the raters scored them using the scoring rubric criteria under the guidance and assistance of the trainer. Although the training session was the main part of the rater training program, it was, however, accompanied by group discussion and score negotiation. These procedures continued until they reached consensus and all raters were confident with determining test takers' scores across the descriptors of the scoring rubrics. Therefore, the phrase "training process" both refers to formal training received in the norming session, and the informal training received through socialization.

The training program consisted of rater training and feedback on previous rating behavior and was conducted in two separate training sessions, each lasting for about six hours, with an interval of one week. Since information about rating behavior was not available until raters completed the first rater training session, the first training session did not include the feedback component. The rater training sessions were held in groups. Individual sessions were also planned to satisfy individual raters' needs during norming and of course to give them feedback on previous rating behavior as part of rater training. In addition to the norming and training sessions, feedback on previous ratings was provided to each rater individually in the second norming session. As Wallace (1991) argued on behalf of second-language rater training, repeated practices do not guarantee development of professional competence. In other words, acquired knowledge can be better internalized with reflection during practice. Thus, according to him, prior rating performance would give raters an opportunity to reflect on their rating behavior. Since each rater had a different rating ability and exhibited various rating behavior, feedback was provided to each rater individually.

The feedback provided to the raters included the raters' previous rating patterns determined by the statistical analysis of the analytic rating scores that each rater gave (i.e. severity, internal consistency, and biases (interaction) with a certain test taker group, task, and rating scale). Regarding feedback on raters' biases, the raters having z-scores beyond ± 2 were considered to have a significant bias and were reminded individually to mind the issue accordingly. With respect to feedback on raters' consistency, the raters having infit mean squares beyond the acceptable range of 0.6 to 1.4, as suggested by Eckes (2015), were considered as misfitting in a way that the raters with an infit mean square value below 0.6 as too consistent (overfit the model) and those with an infit mean square value of above 1.4 as inconsistent (underfit the model). Therefore, the raters were pointed out individually on the issue if they were identified as misfitting.

3.3.3. Post-training phase

Immediately after the training program, the post-training oral test was administered. Having done so, they were provided with the post-training questionnaire to get their attitudes, feelings, their achievements, and evaluations of the training program. Meanwhile, due to the fact that some raters might not have answered all the items of the post-training questionnaire in details, they were also interviewed based on the same questions of the post-training questionnaire to get their ideas and reactions regarding the norming session in detail. All the interviews were audio-taped to have them along with the post-training questionnaires for qualitative data analysis.

3.4. Data analysis

In order to investigate the research question outlined already, the researcher employed a pre-post, mixed-methods research design in which a combination of quantitative and qualitative approaches were used to investigate the raters' development over time with regard to their attitudes in rating L2 speaking performance (Cohen, Manion, & Morrison, 2007). This method offered a comprehensive approach to the investigation of the research questions involving a comparison of raters' perceptions before and after the rater training program.

Qualitative data analysis was done through the analysis of audio-taped interviews, the questionnaires, and the observations of the video recordings of the norming session. Raters' responses to the questionnaires and interviews were analyzed and coded by the researcher. The coding schemes that were employed during data elicitation were based on positive and negative comments using axial coding for the responses given by the raters to the questionnaire items. The analysis of the qualitative data required going through the collected data several times and coding them very precisely. Besides, for the quantitative data analysis, the following analytical approaches were employed: ANOVA to identify any significant difference among the factors identified in the exploratory factor analysis (EFA). Pearson's Product Moment Correlation was used to measure the correlation coefficient among the identified loaded factors of EFA outcomes. EFA was used to analyze and identify the influential factors involved in the raters' attitudes and perceptions both before and after the training program. Confirmatory factor analysis (CFA) was used to neutralize the influential effect of other loading items loaded in each factor and only accounted for the determining items. This was done to get the maximum loading of only those items which have been loaded in each factor at a desired eigenvalue.

4. Results

4.1. Pre-training questionnaire

With regard to *questions number 1 and 2* ($\bar{X}_{Q1} = 0.85$ and ($\bar{X}_{Q2} = 0.9$, respectively) which asked raters about their interest and enthusiasm in engaging a training program, a majority of the raters were positive. Similarly, *question number 3* ($\bar{X}_{Q3} = 0.95$) asked raters about the essentiality of assessing students' oral language proficiency almost and all raters expressed their agreement. Rater OLD5 believed that "It is absolutely important to assess all students' speaking ability so that we will be able to monitor their performance and trace their development path."

Question number 4 ($\bar{X}_{Q4} = 0.65$) interrogated raters about the need for administering formal training programs to achieve desirable rating standards. There were contradictory comments in this respect. A majority of NEW raters and some OLD ones expressed their agreement that having a formal training program maximizes rating quality and standardizes assessment. Rater NEW4 commented that "there is definitely a need to establish training programs to achieve agreement among the raters and develop rating quality among raters." However, in contrast a number of OLD raters believed in reverse and indicated that training programs do not have the desirable effect in standardizing rating among raters. For example, raters OLD8 argued that "I doubt whether training programs could be effective enough in developing raters scoring job considerably."

Question number 5 ($\bar{X}_{Q5}$ = 0.80) asked raters whether they would anticipate any problems in the training program. In response, few raters expressed any problems before commencing the training program; however, rater NEW3 expressed worrisome that "what if after the training program I don't lose difference and get closer to 0, and worse, if I get farther away from 0 after training."

With respect to *question number 6* ($\bar{X}_{Q6}$ = 0.85) which asked raters whether they would predict the training program to be effective, rater NEW9 believed that "I think it will be quite effective if the raters are provided with feedback as well." Rater OLD2 in another comment argued that "it will be effective if it is held regularly."

Questions 7 and 8 ($\bar{X}_{Q7}$ = 0.65 and $\bar{X}_{Q8}$ = 0.65, respectively) dealt with raters' flexibility in accepting the trainer's instructions and to what extent they welcome the trainer's and other authorities' viewpoints. All, except a few OLD raters, confirmed that it is OK for them to notice their mistakes, and that they feel quite comfortable to notice someone else reminds them about their rating errors. Rater NEW9 stated that "I feel quite relaxed and even glad to notice that the trainer spots my mistakes and teaches me something to enhance my rating skill." However, a few OLD raters expressed their dissatisfaction if someone notices their mistakes or tells them something which is against their rating approaches. Rater OLD4 argued that "I don't think I can change my rating approach after the training program. Still, I suppose, I will run my own rating trend."

Question number 9 ($\bar{X}_{Q9}$ = 1.0) asked the raters about their viewpoints of the purpose of the training program. The raters realized its purpose in enhancing raters' accuracy, consistency and bias reduction in rating and they seemed to have been optimistic about such goals.

Questions 10 to 13 ($\bar{X}_{Q10}$ = 0.65, $\bar{X}_{Q11}$ = 0.60, $\bar{X}_{Q12}$ = 0.55, and $\bar{X}_{Q13}$ = 0.65, respectively) asked the raters about their ideas, attitudes, and appropriateness of either the direct or the semi-direct test version for assessing test takers' oral assessment. There were contradictory remarks in this respect. Some raters expressed that due to the interactive nature of speaking skill, it is better to benefit from the direct test version for assessment since it is more realistic to the speaking ability that students need in real contexts. However, some others believed that semi-direct tests are less intimidating for test takers. They added that since the rater is missing in such tests, test takers will be able to reach their highest oral abilities in a stress-free atmosphere. A few also believed that such a test is more feasible to administer.

In order to analyze the raters' responses to the pre-training questionnaire statistically to have a more objective view of raters' perceptions and attitudes of the training program, and to identify how many different aspects of the training program the raters were able to distinguish, an EFA was initially administered. Afterward, an ANOVA, based on the EFA, was run to measure the existence of any significance differences between the raters' responses to the pre-training questionnaire items. Table 2 displays the EFA administered to demonstrate the influential factors related to raters' attitudes and perceptions prior to the administration of the training program. The coding schemes that emerged during data elicitation were classified into various categories including positive and negative comments. It is worthy to note that the scree plot of the eigenvalues which are reported in Table 3 produced an elbow at the third eigenvalue. The first eigenvalue accounted for about 42% of the total variance. According to Henning, Hudson, and Turner (1985) if the loading of the first factor is higher than 20 percent of the whole variance, the required multidimensionality of the data analysis is reiterated. Table 3 displays the result of Principle Axis Factoring at the pre-training phase. As factors of language skills have a high degree of similarity, the direct oblimin was used as the estimation method of rotation in the principle axis factoring of factor extraction.

It is observable from the table of EFA that there were three determining factors loaded with an eigenvalue greater than 0.4 showing that there were three significant factors in raters' viewpoints about the training program prior to its administration. The questionnaire items loaded significantly

Table 2. Exploratory factor analysis of raters' attitudes and perceptions (pre-training)

	Factor		
	1	2	3
1. I feel comfortable with trying a face-to-face rater training program	0.353	**0.812**[a]	−0.196
2. I think I am going to enjoy the rater training experience	0.254	**0.837**[a]	−0.381
3. Generally, I support the notion that we need to assess the language proficiency of students' speaking	**0.860**[a]	−0.305	−0.208
4. It's not necessary to have some sort of formal assessment and training process to ensure comparability of standards	**0.783**[a]	0.328	0.302
5. Do you anticipate any problems with the training program?	0.035	0.136	0.048
6. How effective do you think the rater training program will be?	**0.694**[a]	−0.397	−0.355
7. I am flexible and I do accept authorities' comments in rating even if they are against mine	0.336	−0.313	**0.510**[a]
8. It is difficult for me to notice my mistakes in rating	0.332	0.161	**0.715**[a]

Note: Extraction method: Principal axis factoring.

[a]Three components extracted.

Bold values indicate the factor under which each item of the questionnare has had significant loading.

Table 3. Principle axis factoring of raters' perceptions of the training program (pre-training)

	Initial Eigenvalues		
Factor	**Total**	**% of Variance**	**Cumulative %**
1	2.860	42.127	42.127
2	1.594	19.925	69.451
3	1.103	13.792	83.243
4	0.706	8.831	92.074
5	0.586	6.752	94.334
6	0.350	4.378	96.451
7	0.197	2.461	98.913
8	0.087	1.087	100.000

Note: Extraction method: Principal axis factoring.

in each factor were marked in bold on the table as well. The researcher, then meticulously examined the factor patterns to identify their underlying dimensions on the basis of the highest factor loadings through having a careful attention to the pre-training questionnaire items. The loaded factors were named as the following:

Factor 1 "Enthusiasm of training participation" shows to what extent the raters would like to take part in the training program, and feel comfortable engaging in rater training norming sessions. Item numbers 3, 4, and 6, having the loadings of 0.69 and above, were loaded in this factor.

Factor 2 "Efficacy of the training program" shows to what extent raters predict the effectiveness of training program in reducing rater variation (severity/leniency and biasedness) and increasing internal consistency. Besides, whether they feel like the need for training programs to reach the above-mentioned goals. Item numbers 1 and 6, having the loadings of 0.81 and above, were loaded in this factor.

Factor 3 "Rater flexibility" demonstrates to what extent raters are flexible and willing enough to accept the authorities' comments and whether they are open to reality and welcome noticing their mistakes in rating. Item numbers 7 and 6, having the loadings of 0.51 and above, were loaded in this factor.

It is must be indicated that item number 5, asking the raters whether they predicted any problems in the training program, was not loaded in any factor. This shows that the raters had a contrasting viewpoint in this respect in a way that some predicted problems occurring during the norming session, whereas the rest did not. Afterward, the scores for the three factors were then used as dependent variables in an ANOVA test to identify whether there was a significant difference among the raters' attitudes and perceptions of the training program prior to its administration. Table 4 displays the ANOVA analysis of raters' attitudes and perceptions of the training program prior to its administration.

The finding shows that there was a significant difference with respect to raters' responses to the pre-training questionnaire items as obtained by factor analysis. The drawback of the EFA is that the identification of the influential factors is seen based on maximum loading amount of items under each factor. However, this, by no means indicates that the other items do not have any effect on the loading of the item(s) loaded under a factor. In other word, in EFA, the loading effect of other items have not been neutralized but simply ignored. It must be argued that, however, all the other test/questionnaire items have their own loading effect on any particular item(s) loaded on a particular factor as well. Accordingly, in order to only evaluate the loading effect of the item(s) loaded on each factor and thus to neutralize the loading effect of all other items, a CFA was run. CFA was performed using IBM Amos version 22.0. Table 5 displays the CFA administered to demonstrate the influential factors related to raters' attitudes and perceptions prior to the administration of the training program. The model fit indices of CFI, TLI, RMSEA, and SRMR display that the model used to obtain raters' attitudes and perceptions of the training program, at the pre-training phase, was a good and suitable model.

Table 4. ANOVA analysis of factor scores of raters' perceptions of the training program (pre-training)

	Sum of squares	*df*	Mean square	*F*	Sig.
Between groups	78.844	2	39.422	4.842	0.019
Within groups	170.965	21	8.141		
Total	249.809	23			

Table 5. Confirmatory factor analysis of raters' attitudes and perceptions (pre-training)

	Factor		
	1	2	3
1. I feel comfortable with trying a face-to-face rater training program		0.851	
2. I think I am going to enjoy the rater training experience		0.896	
3. Generally, I support the notion that we need to assess the language proficiency of students' speaking	0.947		
4. It's not necessary to have some sort of formal assessment and training process to ensure comparability of standards	0.731		
5. Do you anticipate any problems with the training program?			
6. How effective do you think the rater training program will be?	0.801		
7. I am flexible and I do accept authorities' comments in rating even if they are against mine			0.634
8. It is difficult for me to notice my mistakes in rating			0.765

CFI: 0.893.

TLI: 0.886.

RMSEA: 0.591.

SRMR: 0.790.

Table 6. Interfactor correlation coefficient of the raters' perceptions of the training program (pre-training)

		F1	F2	F3
F1	Pearson Correlation	1	0.077	0.001
	Sig. (2-tailed)		0.856	0.998
	N	8	8	8
F2	Pearson Correlation	0.077	1	0.138
	Sig. (2-tailed)	0.856		0.745
	N	8	8	8
F3	Pearson Correlation	0.001	0.138	1
	Sig. (2-tailed)	0.998	0.745	
	N	8	8	8

It is noteworthy to indicate through the analysis of raters' responses to the questionnaire items, it was found that some OLD raters were pessimistic and did not have a positive attitude with respect to the usefulness and effectiveness of the training program in reducing rater variation (severity/leniency and biasedness) and increasing internal consistency among them. However, almost all NEW raters were quite positive in this respect. For the rest of the items, there were no substantial differences in between the responses of the two groups of raters as long as the coding of qualitative data provided the researcher with.

Interfactor correlations for the raters' responses at the pre-training phase were measured using the Pearson's Product Moment Correlation in order to determine which factors have the highest go togetherness with one another. Table 6 represents the correlation coefficient of the determining factors as identified by CFA. The highest correlation coefficient was found between factor two (efficacy of the training program) and factor three (rater flexibility) measured 0.13, $p > 0.05$. However, no significant relationship, as the outcome displays, was observed between them. The lowest correlation coefficient was found between factor one (enthusiasm of training participation) and factor three (rater flexibility) measured 0.001, $p > 0.05$.

4.2. Post-training questionnaire

The *first question* ($\bar{X}_{Q1} = 0.85$) dealt with the purpose of the training program and asked the raters whether the training program was effective enough or not. Raters were rather mostly positive. For example, rater NEW5 stated that "receiving feedback from the trainer and other raters was very helpful." Rater OLD10 indicated that "receiving comments on rating was very constructive." However, still some raters expressed apprehension in this respect, specifically on their own side. For instance, rater NEW3 expressed that "I'm not sure if I've gained enough of the training program and if it was really effective on me." Rater OLD5 also noted that "I doubt whether I'm going to act as what I was trained for in the training program."

Questions two to five ($\bar{X}_{Q2} = 0.95$, $\bar{X}_{Q3} = 0.90$, $\bar{X}_{Q4} = 1.0$, and $\bar{X}_{Q5} = 0.90$, respectively) mostly asked raters about the trainer's behavioral characteristics, whether or not they felt intimidated talking to him and whether or not they understood what was instructed by the trainer during the norming sessions. All the raters in both expertise groups were positive and expressed that they could freely and without any anxiety talk to the trainer. In this respect, rater NEW7 stated that "The trainer was so friendly and knowledgeable and I felt quite relaxed interacting with him in a stress-free atmosphere."

The *sixth question* ($\bar{X}_{Q6} = 0.85$) in the post-training questionnaire asked the raters whether the enjoyed the training program or not. Although a majority of the raters responded "Yes" to the question, rater OLD4 found it rather *tiring*. A number of NEW raters stated that although they benefited

from the training program, they were rather dominated by outspoken and more frank raters and due to their shyness they could not manifest their capabilities (Raters NEW2, NEW5, and NEW6).

Questions seven and eight ($\bar{X}_{Q7} = 0.85$ and $\bar{X}_{Q8} = 0.90$, respectively) dealt with raters' own judgments of the success or failure of the training program and the teaching materials (including, texts, sample tests, audio and video materials ...) used for training. In this respect, a majority of the raters expressed that the program was quite successful and it helped them enhance their rating skill to a great extent. Rater NEW2 believed that "the training program was truly beneficial for me. It revolutionarily changed my view about rating." Rater OLD1 also argued that "the training program was a good practice to develop rating. It opened a wider gate of rating across from my eyes, and provided me with various techniques and strategies to rate." However, it must be reiterated that a few raters doubted about the successful outcome of the training program in developing raters' rating. They further added that the training program little change their rating behaviors.

The *ninth question* ($\bar{X}_{Q9} = 0.85$) asked the raters whether they noticed any change in their rating performance following the training program. The responses were various. Some raters stated that the training program changed their rating quite drastically. Rater NEW3 stated that "training enabled me to do the rating job with a lot more self-confidence and I felt a lot more relaxed." However, some others, specifically those who had already been negative about the training program in the pre-training questionnaire, argued that their ratings changed just a little compared to before and that the training program had little success (Raters OLD8 and OLD7). There also seemed to be a sort of halo effect in which the raters awarded similar scores to test takers on the categories of the rating scale. In this respect rater OLD6 stated that "I just gave this score to her cohesion and intelligibility. So, I think I am going to give the same score to her grammar and vocabulary."

The *tenth question* ($\bar{X}_{Q10} = 0.85$) asked the raters about the feedback provided during the training program for raters on their ratings before the training program. A majority of the raters found the feedback quite useful and believed that it had a constructive influence on their scoring behavior. Rater NEW9 stated that "It was good to see where I had got farther from the rating of other raters." Rater OLD2 commented that "the feedback made me think about what I had done and was a good reflection on what I had done." Some raters, including raters OLD8 and OLD4 questioned the effectiveness of the given feedback and they argued that the feedback changed their rating behavior just very little.

It must be indicated that raters' responses to the feedback question had a close relationship with their receptiveness of the feedback and thus the occurrence of change in their rating behaviors. For example, Rater NEW8 had a positive attitude to the provision of feedback prior to the training program. Consequently, after the training program she found it quite useful and stated that "training program helped me better match my rating with the ability of the test takers." This rater was significantly lenient at the pre-training phase and could reduce leniency to a high extent. As another example, rater OLD5 was significantly sever prior to the training program and also optimistic about the feedback to be provided for the raters. After training, he found the feedback very effective and was able to modify his rating behavior, according to the feedback, to a considerable extent thus bringing him to an acceptable range of biasedness. For most of the other raters, the result was quite successful and raters indicated that they benefitted quite a lot from the training program. This had also been reflected in the quantitative statistical data analysis using FACETS which demonstrated that they achieved more consistency after training. A few raters, e.g. rater OLD8, who although had enhanced consistency (moving from InfitMnSq. of 0.3 to 0.5) after training, were still away from acceptable range. These raters were all among those who were pessimistic about the training program and feedback and found it little or no effective for them. In this respect, rater OLD8 commented that "the feedback was no useful at all and it didn't change my rating." Similarly, rater OLD4 argued that "I'm not sure whether my rating has changed any compared to before." Such finding tells us that we can trace a one-to-one relationship between raters' attitude and perception about the training program and feedback, before training and its effect on them in efficacy power.

The *eleventh question* asked the raters about the benefits of the training program. A substantial number of the raters stated that they had the advantage of having interaction with other raters and sharing their ideas and understandings of the training program with each other. Rater NEW7 believed that "this made me have the opportunity to be aware of other raters' viewpoints and get help in case of necessity. It was fun and amusing." In another commentary, rater OLD5 stated "I really loved the friendly atmosphere of the work. Being a part of a team that everyone else is working on the same thing is quite fun and enjoyable." Rater NEW7 argued that "the training program was a good practice to compare performances and it was fun." Rater NEW9 stated that "receiving feedback was very helpful to let the raters notice whether or not they have done what was intended by the training program."

The *twelfth question* asked the raters about the disadvantages of the training program. In this regard, rater NEW6 stated that "the problem is that some raters cannot cope with their own time and pace of rating." In a further commentary rater OLD2 believed that "it was rater getting tiring at times and raters were not able to take breaks when they felt tired." Raters NEW1 indicated that "being worried about what other raters might have assigned to test takers' performances made me rather unable to concentrate on my work."

Similar to the pre-training phase, in order to analyze raters' responses to the post-training questionnaire, an EFA was run to determine the influential factors having a significant role in identification of raters' perceptions and attitudes following the training program. Afterward, an ANOVA, based on the outcome of the EFA, was run to measure the existence of any significance differences between the raters' responses to the post-training questionnaire items. Table 7 displays the EFA administered to demonstrate the influential factors related to raters' attitudes and perceptions following the administration of the training program. At this phase, the scree plot of the eigenvalues reported in Table 8 produced an elbow once again at the third eigenvalue. The first eigenvalue accounted for about 46% of the total variance. Table 8 displays the result of Principle Axis Factoring at the post-training phase.

Similar to the pre-training phase, three determining factors were loaded with an Eigenvalue greater than 0.4 showing that there were three significant factors in raters' viewpoints following the administration of the training program. The questionnaire items loaded in each factor were marked in bold on the table as well. The loaded factors were named as the following:

Table 7. Exploratory factor analysis of raters' attitudes and perceptions (post-training)

	Factor		
	1	2	3
1. Altogether how effective did you find the face-to-face training program?	**0.876**[a]	−0.159	−0.262
2. Overall, how friendly (trainer-trainee) did you find the face-to-face training program?	0.339	**0.778**[a]	0.268
3. The rater trainer was tense	0.377	**0.655**[a]	−0.262
4. I could talk to the trainer easily	0.327	**0.591**[a]	0.291
5. The trainer used understandable language	**0.788**[a]	0.188	−0.163
6. How much did you enjoy your face-to-face training experience?	0.331	−0.387	**0.469**[a]
7. How much do you think the face-to-face training program achieved its purpose?	**0.882**[a]	−0.024	−0.144
8. How much material descriptors, (e.g. scripts, notes, etc.) was used in the program?	**0.693**[a]	−0.518	−0.236
9. How much, do you think, your method of rating changed as a result of face-to-face training program?	**0.670**[a]	−0.510	−0.010
10. How effective did you find the feedback given to you during the norming session?	**0.516**[a]	0.187	0.204

Note: Extraction Method: Principal Axis Factoring.

[a]Three components extracted.

Bold values indicate the factor under which each item of the questionnare has had significant loading.

Table 8. Principle axis factoring of raters' perceptions of the training program (post-training)

Factor	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.930	46.306	46.306
2	1.545	15.455	64.755
3	1.188	11.878	76.633
4	0.704	7.035	83.668
5	0.480	4.804	88.472
6	0.397	3.971	92.443
7	0.308	3.083	95.526
8	0.208	2.079	97.606
9	0.130	1.296	98.902
10	0.110	1.098	100.000

Note: Extraction Method: Principal Axis Factoring.

Factor 1 "Effectiveness of the training program" shows to what extent the raters felt that the training program achieved its purposes and was effective enough in bringing the raters into higher consistency with each other. Also, whether or not the materials used in the study were appropriate enough to assist the raters modify their rating behavior. Item numbers 1, 5, 7, 8, 9, and 10, having the loadings of 0.51 and above, were loaded in this factor. This shows that a majority of the raters felt quite positive about the constructive effectiveness of the training program on raters in bringing them into more consistency and reducing bias effect.

Factor 2 "Stress-free atmosphere" shows to what extent raters felt relaxed interacting with the rater trainer and whether or not the training program was stress-free for them and that the raters were comfortable during the training program. Item numbers 2, 3, and 4, having the loadings of 0.59 and above, were loaded in this factor showing that the raters felt comfortable attending the training program.

Factor 3 "training enjoyment" demonstrates whether or not and to what extent raters enjoyed taking part in the training program. Item 6, having the loading of 0.46, was loaded in this factor reflecting that the raters enjoyed the training program.

Afterward, similar to the pre-training phase, the scores for the three factors were then used as dependent variables in an ANOVA test to identify whether there was a significant difference among the raters' attitudes and perceptions following the training program. Table 9 displays the ANOVA analysis of raters' attitudes and perceptions following the administration of the training program.

The finding showed that there was a significant difference with respect to raters' responses to the post-training questionnaire items as obtained by EFA. As already mentioned regarding the shortcoming of the administration of EFA in determining the influential factors of raters' perceptions of the training program, in order to only evaluate the loading effect of the item(s) loaded on each factor meanwhile neutralizing the loading effect of all other items, a CFA was run. Table 10

Table 9. ANOVA analysis of factor scores of raters' perceptions of the training program (post-training)

	Sum of squares	*df*	Mean square	*F*	Sig.
Between groups	1.877	2	0.938	7.114	0.004
Within groups	3.165	27	0.132		
Total	5.042	29			

Table 10. Confirmatory factor analysis of raters' attitudes and perceptions (post-training)

	Factor		
	1	2	3
1. Altogether how effective did you find the face-to-face training program?	0.908		
2. Overall, how friendly (trainer-trainee) did you find the face-to-face training program?		0.851	
3. The rater trainer was tense		0.688	
4. I could talk to the trainer easily		0.613	
5. The trainer used understandable language	0.741		
6. How much did you enjoy your face-to-face training experience?			0.574
7. How much do you think the face-to-face training program achieved its purpose?	0.819		
8. How much material descriptors, (e.g. scripts, notes, etc.) was used in the program?	0.723		
9. How much, do you think, your method of rating changed as a result of face-to-face training program?	0.746		
10. How effective did you find the feedback given to you during the norming session?	0.552		

CFI: 0.911.

TLI: 0.906.

RMSEA: 0.579.

SRMR: 0.781.

demonstrates the CFA results demonstrating the influential factors related to raters' attitudes and perceptions following the administration of the training program. The model fit indices of CFI, TLI, RMSEA, and SRMR displayed that the model used to obtain raters' attitudes and perceptions of the training program, at the post-training phase, was a good and suitable model.

A further complementary analysis of the raters' responses to the post-training questionnaire items indicated that a majority of the raters expressed that they have benefitted quite a lot from the training program and stated that the training program was useful in bringing more consistency and reducing severity/leniency and bias effect in their ratings. However, those few OLD raters who had already expressed their pessimistic view of the training program, at the pre-training phase, again indicated that they gained little after training. It is noteworthy to indicate that these raters were the ones who did not improve or improved very little with respect to the reduction of severity/leniency and bias indices and achieved more consistency after the training program.

Interfactor correlations for the raters' responses at the post-training phase were measured using the Pearson's Product Moment Correlation in order to determine which factors have the highest go togetherness with one another. Table 11 represents the correlation coefficient of the determining factors as identified by CFA. The highest correlation coefficient was found between factor one (effectiveness of the training program) and factor two (stress-free atmosphere) measured 0.83, $p < 0.01$. However, contrary to the pre-training phase, this time a significant correlation was found. The obtained correlation between the two above-mentioned factors was found much larger than typical according to Cohen's table of effect size. The lowest correlation coefficient was found between factor two (stress-free atmosphere) and factor three (training enjoyment) measured 0.365, $p > 0.05$.

A majority of the raters (14 raters) claimed that the feedback was not only easy to understand but also they were able to use it in their own ratings. Three of them even reiterated that the provision of feedback even assisted them shed more light on their rating behaviors for example, rater OLD5 said that "the feedback and the training program always will be kept in my mind when rating an oral performance." However, two raters OLD8 and OLD4 questioned the effectiveness of the training program. Rater OLD4 argued that "the training program and the feedback did not have any effect on my ratings." And rater OLD8 reported that "I suppose the training program even made me severer

Table 11. Interfactor correlation coefficient of the raters' perceptions of the training program (post-training)

		F1	F2	F3
F1	Pearson Correlation	1	0.839**	0.675
	Sig. (2-tailed)		0.009	0.066
	N	10	10	10
F2	Pearson Correlation	0.839**	1	0.365
	Sig. (2-tailed)	0.009		0.374
	N	10	10	10
F3	Pearson Correlation	0.675	0.365	1
	Sig. (2-tailed)	0.066	0.374	
	N	10	10	10

**Correlation is significant at the 0.01 level (2-tailed).

Table 12. Relationship between raters' questionnaire and interview responses and their rating behavior change

	Severity		Bias		Consistency	
	N	**%**	***N***	**%**	***N***	**%**
Feedback successful/positive about feedback	13	65	14	70	14	70
Feedback successful/negative about feedback	0	0	3	15	5	25
Feedback unsuccessful/positive about feedback	0	0	0	0	0	0
Feedback unsuccessful/negative about feedback	7	35	3	15	1	5

in my ratings." Regarding the issue of consistency, as a part of feedback provision, a majority of the raters (14 raters) responded positively; however, rater NEW6 felt confused over the concept. He said "It was difficult for me to understand the feedback for consistency. Honestly speaking, I didn't know what consistency meant. I think I did much better on biasedness and severity level after training, but perhaps not about consistency."

Table 12 displays that there was a positive relationship between their perceptions of the feedback and the receptiveness and success of the feedback for a majority of the raters. This was done by having the raters score the test takers' oral performances both before and after the training program. The collected data were analyzed using the multifaceted Rasch measurement (MFRM) to identify their measures of consistency, severity, bias in the two phases of the study. Since the analyses of the MFRM are beyond the scope of this article, the importance of this is to compare and contrast raters' positive or negative perceptions about the training program and the effectiveness of feedback provision (the current study) and the reduction of their biases and severity and increase in their consistency measures after training. The outcome indicated that 65 percent of the raters in severity, 85 percent in bias, and 95 percent in consistency fell into the first two table rows indicating the success of the training program and feedback provision but having either positive or negative view about it. Thirty-five and 15 percent of them in severity and bias and only 5 percent of raters in consistency were reported not to have benefited from the training program and feedback provision.

In sum, the raters were generally positive about the training program and the provision of feedback on the former ratings. Although, there were some slight hesitations about the complete incorporation of the training principles in the subsequent ratings, a majority of them felt quite satisfied with the training program and the feedback provided for them.

5. Discussion

The results of this study are different from the previous research findings in that this piece of research has also been unique in that it regarded the application of feedback with respect to oral proficiency assessment to investigate any possible relationship between raters' perceptions and attitudes of the training program and feedback on their real change of rating behavior. The outcome of this study can be classified into two parts: First, it was found that raters' perceptions of the feedback and the training program were generally positive and promising. A majority of them were positive about the program and thus improved their rating quality. They mostly agreed that the feedback helped them think more carefully and use logic about what they were doing. After receiving feedback, most raters modified their rating behavior regarding severity/leniency and biasedness and enhanced consistency for which both the qualitative and quantitative data analysis were absolutely in-line with each other.

Three raters (OLD8, OLD4, and OLD7) were still extremely biased after training and rater OLD8 was consistent. These raters were the ones who expressed negativity about the training program beforehand. It is noteworthy to indicate that, even those who were not optimistic about the program or even intentionally ignored it could all modify their consistency and in one case a rater could improve his severity and bias as well and brought it within the acceptable range. It is quite interesting to indicate that no raters with a positive attitude on feedback received a negative result. This finding is in contrast with that of Knoch (2011) who found in her study little relationship between raters' perceptions and the usefulness of the feedback and the actual success of the feedback. She found, for a majority of the cases, a mismatch between the success of the feedback and whether the raters thought they had successfully included it in their ratings. However, the finding is parallel with the one found by Nakatsuhara (2011) and Van Moere (2012) who in separate studies observed a high correlation between raters' perceptions of training programs and the impact of feedback on their subsequent ratings.

Data analysis indicated that through rater training programs, rater effects, and rater variability, as pointed out by In'nami and Koizumi (2016), could be controlled and neutralized. This will result in less halo effect which will provide decision-makers with reduced levels of bias and severity and increase in consistency in measurement. The outcome of this study demonstrated that although the raters claimed that they underwent some difficulty including shyness and lack of confidence attending the training program, it was quite useful in reducing raters' severity and biases and increasing their consistency level. Most particularly, because the raters were provided with post-training individual and group feedback, they were able to reduce their biases quite magnificently. This finding is fairly consistent with that of Knoch et al. (2007) who found that raters reduced biases more in a face-to-face training program, compared to the online rater training, since they were given feedback after rating.

Second, in spite of the reduction of raters' severity/leniency differences after the training program, there still observed considerable amount of variation among the raters. This weird phenomenon was because of the extremist behavior of a number of raters including OLD8, OLD4, OLD7 (in severity), and OLD3, OLD9, and NEW6 (in leniency) who had already expressed their mistrust in the effectiveness of the training program, the reason of which could be their arrogance, overconfidence or pessimism. Consequently, that is why they almost did not or very little improved their rating behavior after the training program. Therefore, the overall performance of the raters was affected by this causing significant variation even after training. This outcome is consistent with that of Bijani (2010) and Davis (2016) who found the impact of training programs more important than raters' expertise.

It can be hypothesized that those raters whose rating behavior improved rather little after the training program were among those who had a negative attitude toward the training program and its effectiveness in reducing raters' biasedness and increasing consistency. This result is parallel with Hamilton et al.'s (2001) and Kuiken and Vedder (2014) finding who found a relative linear

relationship between raters' perceptions of training program and its effectiveness on them. Meanwhile, it is noteworthy to indicate that although a one-to-one causal relationship between the raters' attitudes and their rating performance cannot be drawn, it can be presumed that if training programs can satisfy raters' expectations, they will lead to a higher consistency among raters and the benchmark accordingly. So on the first step, it seems rather necessary to build enough interest among raters in the training programs they are supposed to undergo.

The outcomes of this research showed that raters were required to be aware of the goal of performance assessment in the training program. With this, rating would result in less halo effect. This is in line with what Attali (2016) found in his study on the assessment of test takers' oral performance through training raters. The study also showed that the raters with more positive attitudes toward rating feedback were able to incorporate it more successfully in their rating and thus achieved more consistency and less biasedness in their subsequent ratings. Similar to the above finding, the study showed that although a one-to-one relationship between the raters' attitudes and their rating performance cannot be formulated, it can be assumed that if training programs can satisfy raters' expectations, they will lead to a higher consistency among raters and the benchmark accordingly. In other words, training program was generally more useful for those raters who were optimistic and responded positively regarding its effective outcomes. This is parallel with Davis (2016) finding emphasizing the greater impact of training on the raters having a more optimistic view about it. One further very important implication of this study is on behalf of the large number of participants taking part in this study. Since the findings of this study were derived from a pool of 300 test takers, the reliability measures and qualitative validity obtained in this study can be reported with higher certainty. This definitely makes the generalization of the findings of the study less risky.

One very considerable finding of this study regarding bias interaction between raters and test takers was that a higher bias interaction was observed for test takers on the extreme high continuums of language performance ability, i.e. high ability test takers. This finding is both similar to and different from that of Winke et al. (2012) and Kim (2015) who found that raters showed extreme bias toward both extreme continuums of test takers' oral ability performance, i.e. both high-ability and low-ability test takers. This finding might be due to that fact that some raters' expectations of test takers would rise when rating higher ability test takers than lower ability test takers although the finding was not as considerable as it was for higher ability ones to benefit them to get the passing mark. Besides, this finding calls for a need of clearer rating criteria and a more extensive and comprehensive training program with more significant focus on rating test takers on the extreme ability points.

A possible reason for observing contradictory findings between this study and the previous studies (e.g. Kenyon & Stansfield, 1991; Nakatsuhara, 2011; Winke et al., 2012) could be attributed to the use of different raters in different assessment settings and different test items and tasks. Raters' various perceptions could also be a determining issue in this respect as well. In another point of view, the findings of this study, regarding raters' biases toward test takers, is quite useful in a way that in Iran evaluation is typically norm-referenced which means that rating is done through making comparisons with the performances of other test takers (Farhady & Hedayati, 2009). Thus, the findings could provide testing centers and decision-makers with better strategies to improve the fairness of scoring. However, since this study was done on a particular group of raters and test takers and in a specific context, further research is required to provide more evidence to confirm the above-mentioned findings.

6. Conclusions

The analysis of the norming sessions showed that some of the criteria in the scoring rubric were rather vague for inexperienced raters to use, thus they should be instructed to the raters orally during the training program. The outcome of the study showed that raters with a positive attitude about training programs will benefit more from the provision of feedback. In other words, training programs and feedbacks are more influential for the raters who are optimistic about them. Likewise, for

those raters who felt negative about training programs, the impact reduced to the minimum. Moreover, raters' background, specifically their level of experienced was shown to have, on average, negative impact on the effectiveness of the training program. Presumably, due to the arrogance or overconfidence of experienced raters, they did not enjoy as much improvement in rating quality as inexperienced raters did with regard to the factors of severity, bias, and consistency. The implication is that decision-makers had better not be concerned about raters' expertise to enhance the quality of oral performance assessment. Since experienced raters demand higher budgets for rating, decision-makers had better invest their budgets on establishing appropriate training programs. Raters' positive perceptions and attitudes about training programs resulted in less halo effect and decreased their variability in assessing test takers' performance abilities. Consequently, since through rater training programs rater effects and variability can be controlled, decision-makers had better establish rater training programs to increase rater consistency and reduce their biases in measurement.

The questionnaires and interviews used in this study were the modified versions of ready-made ones which were adapted on the basis of the context of the research. Similar studies could be done using researcher-developed questionnaire and interviews in order to compare and contrast raters' and test takers' perceptions in various contexts. More research could be run investigating the influence of raters' background and personality related issues (e.g. various L1 backgrounds and accents) on raters' consistency of scoring and new trends in running training programs accordingly.

Funding
The author received no direct funding for this research.

Author details
Houman Bijani[1]
E-mail: houman.bijani@gmail.com
[1] Department of English Language Teaching, Islamic Azad University, Zanjan, Iran.

Citation information
Cite this article as: Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training, Houman Bijani, *Cogent Education* (2018), 5: 1460901.

References

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99–115. https://doi.org/10.1177/0265532215582283
Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49–58.
Bijani, H. (2010). Raters' perception and expertise in evaluating second language compositions. *The Journal of Applied Linguistics, 3*(2), 69–89.
Bijani, H., & Fahim, M. (2011). The effects of rater training on raters' severity and bias analysis in second language writing. *Iranian Journal of Language Testing, 1*(1), 1–16.
Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1–25. https://doi.org/10.1191/0265532203lt242oa
Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16–33. https://doi.org/10.1177/026553229501200102
Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. London: Routledge.
Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117–135. https://doi.org/10.1177/0265532215582282
Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies*. Oxford: Oxford University Press.
Eckes, T. (2015). *Introduction to many-facet Rasch measurement*. Frankfurt: Peter Lang Edition.
Elder, C., Barkhuizen, G., Knoch, U., & von 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
ETS (2014). *ETS oral proficiency testing manual*. Princeton, NJ: Author.
Farhady, H., & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics, 29*(1), 132–141. https://doi.org/10.1017/S0267190509090114
Fulcher, G. (2003). *Testing second language speaking*. London: Longman.
Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' perceptions of on-line rater training and monitoring. *System, 29*(4), 505–520. https://doi.org/10.1016/S0346-251X(01)00036-7
Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing, 2*(2), 141–154. https://doi.org/10.1177/026553228500200203
In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing, 33*(3), 341–366. https://doi.org/10.1177/0265532215587390
Kenyon, D. M., & Stansfield, C. W.(1991). *A method for improving tasks on performance assessments through field testing*. Paper presented at the Annual meeting of the National Council on Measurement in Education. Chicago, IL.
Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing, 34*(1), 23–48. https://doi.org/10.1177/0265532215595666
Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly, 12*(3), 239–261. https://doi.org/10.1080/15434303.2015.1049353
Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – A longitudinal study. *Language Testing, 28*(2), 179–200. https://doi.org/10.1177/0265532210384252

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26–43. https://doi.org/10.1016/j.asw.2007.04.001

Kuiken, F., & Vedder, I. (2014). Raters' decisions, rating procedures and rating scales. *Language Testing, 31*(3), 279–284. https://doi.org/10.1177/0265532214526179

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing, 33*(3), 319–340. https://doi.org/10.1177/0265532215587391

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

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*(4), 479–499. https://doi.org/10.1177/0265532214530699

McNamara, T., & Adams, R. (1991). *Exploring rater behavior with Rasch techniques* (ERIC Document Reproduction Service No. ED345498, pp. 1–28). Australian Council for Educational Research.

McNamara, T. F. (1995). Modelling performance: opening Pandora's box. *Applied Linguistics, 16*(2), 159–179. https://doi.org/10.1093/applin/16.2.159

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

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*(4), 483–508. https://doi.org/10.1177/0265532211398110

Overall, J. E., & Magee, K. N. (1992). Estimating individual rater probabilities. *Applied Psychological Measurement, 16*(1), 77–85. https://doi.org/10.1177/014662169201600109

Reed, D. J., & Cohen, A. D. (2001). Revisiting rater and ratings in oral language assessment. In T. McNamara, K. O'Loughlin, C. Elder, & A. Brown (Eds.), *Studies in language testing: Experimenting with uncertainty: Essays in honor of Allan Davies* (pp. 82–96). Cambridge: Cambridge University Press.

Tashakkori, A. & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325–344. https://doi.org/10.1177/0265532211424478

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

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.

Cogent Education **(ISSN: 2331-186X) is published by Cogent OA, part of Taylor & Francis Group.**