Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 11, No. 2, October 2021
ISSN 2476-5880

Fairness in Oral Language Assessment: Training Raters and Considering Examinees' Expectations

Mehdi Doosti[1]*, Mohammad Ahmadi Safa[2]

Received: 18 May 2021 Accepted: 25 June 2021

Abstract

This study examined the effect of rater training on promoting inter-rater reliability in oral language assessment. It also investigated whether rater training and the consideration of the examinees' expectations by the examiners have any effect on test-takers' perceptions of being fairly evaluated. To this end, four raters scored 31 Iranian intermediate EFL learners' oral performance on the speaking module of the IELTS in two stages (i.e. pre- and post-training stage). Furthermore, following Kunnan's (2004) Test Fairness Framework, a questionnaire on fairness in oral language assessment was developed, and after pilot testing and validating, it was administered to the examinees at both stages. The examinees' expectations were taken into account in the second round of the speaking test. The results indicated that rater training is likely to promote inter-rater reliability and, in turn, enhances the fairness of the decisions made based on the test scores. It was also concluded that considering students' expectations of a fair test would improve their overall perceptions of being fairly evaluated. The results of this study sought to provide second language teachers, oral test developers, and oral examiners and raters with useful insights into addressing fairness-related issues in oral assessment.

Keywords: Inter-rater reliability; Oral language assessment; Rater training; Test fairness.

1. Theoretical background and review of the related literature

1.1. Performance-based assessment of language

The standardized testing methods such as multiple-choice, matching, true/false, etc., which are widely used in traditional assessments of student performance, do not seem to be strongly efficient in the direct measurement of their progress (Bland & Gareis, 2018; Luongo-Orlando, 2003). Performance assessment, however, as a complementary alternative to traditional methods of assessment, is designed to directly observe students performing in simulated or real-life contexts. Emphasizing the importance of performance-based assessment in L2 oral testing, Chalhoub-Deville (1995) asserted that "performance-based tests require students to produce complex responses integrating various skills and knowledge and to apply their target language

[1] Kermanshah University of Medical Sciences. mehdi.doosti@kums.ac.ir
[2] Bu_Ali Sina University (Hamedan). m.ahmadisafa@basu.ac.ir

 ISSN 2476-5880

skills to life-like situations" (p. 16). Furthermore, the content and contextual features of performance tasks should represent the situations in which examinees encounter in the real world. Situations like answering the telephone, carrying out university routine activities and studies, going shopping, and communication activities which require examinees to demonstrate their second language proficiency in actual or simulated settings (Wesche, 1987).

Brown (2004) believes that performance-based language assessments are more advantageous than the standardized testing methods in that they assess individuals' performance in simulated or real-world tasks. Students can demonstrate their knowledge, skills, and perceptions through some salient modes of performance assessment procedures such as oral presentations, group performances, written production, projects, and other interactive tasks (Brown, 2004; Conn, Bohanb, Pieper, & Musumeci, 2020). Reviewing extant literature, Bland and Gareis (2018) scrutinized the various definitions of performance assessments and offered the following based on the frequency of the words and concepts used to define performance assessment in the literature:

> Performance assessments, which can develop as a task or product, necessitate subjective judgment to measure students' abilities to authentically demonstrate knowledge, skills, and processes in a way that provides value, interest, and motivation to students beyond the actual score or grade (p. 66).

In this regard, they suggested that there was an urgent need to move from the reliance on standardized, fixed-choice assessments to the use of more authentic assessments of real-world problem-solving and higher-order thinking skills.

However, despite the merits, one of the greatest challenges to the applicability of performance-based assessment has been the concern over the unreliability of the obtained test scores. The concern has been of grave importance and relevance as the subjectivity of the scoring procedure and the lack of detailed and practical scales have been two afflicting problems in the field and a large number of studies have aimed to present rating scales and scoring criteria for both oral and written performances (e.g., Bland & Gareis, 2018; Brown & Bailey, 1984; Chen, 2016; Norris, Brown, Hudson, & Bonk, 2002). On the other hand, studies have also focused on the effects of rater characteristics and bias on language learners' performance (e.g. Caban, 2003; Davis, 2016; Henning, 1996; Kang, Rubin, & Kermad, 2019; Lumley & McNamara, 1995; Wind & Peterson, 2018) as some other threats to the reliability of the scores gained through a performance-based assessment. As Chalhoub-Deville (1995) argued, test scores might be influenced by a test method used to measure a particular construct as well as various rater characteristics such as background characteristics (e.g., professional experiences), standards for performance, and criteria for assessing performance and it is based on such grounds that speaking tests often include scoring rubrics and criteria to guide raters in evaluating an individual's performance (Douglas, 1994; Kozaki, 2010).

Regarding rating scales to assess students' oral performance, for instance, Chen (2016) developed a model of analytic rating scales to assess L2 Chinese oral performance. She defined a rating scale as "a scoring guide used to assess performance against a set of criteria" (p. 51) maintaining that in oral performance assessment, the purpose of the test determines the type of analytic rating scales. In terms of assessing oral performance of second language learners, the

 ISSN 2476-5880

various types of analytic rating scales in the literature she reviewed included accent, pronunciation, fluency, grammatical accuracy, vocabulary, comprehension, and sociolinguistics. However, in Chen's (2016) model, four factors were identified to explain the analytic rating scales in her study: fluency, conceptual understanding, communication clarity, and communication appropriateness, which explained only 65.5% of teachers' holistic judgments of oral performance. Accordingly, she suggested that teachers and assessment professionals add other facets such as students' self-descriptions of their own language ability or longitudinal records of teachers' ratings of students to evaluate oral performance.

1.2. Rater training and its effect on inter-rater reliability

Most testing professionals associate fairness with validity (Amirian, Ghonsooli, & Amirian, 2020) particularly the consequential validity proposed by Messick (1989), and reliability is sometimes considered as a requirement for validity (Bachman, 1990); however, reliability is as significant as validity when discussing fairness in assessment. As Wind and Peterson (2018) put it, construct-irrelevant effects such as demographic variables, characteristics of students, raters, and assessment context, and rater's previous experiences or levels of training are factors considered when discussing fairness in the context of rater-mediated assessment systems. In subjective assessment of performance tests, variance due to rater differences and disposition is a source of error (Kang, Rubin, & Kermad, 2019). According to Bachman (1990), raters judge an individual's language performance based on a set of criteria that define what constitutes an adequate performance. Consistent and invariable application of the same set of criteria will produce a reliable set of ratings. In the case of ratings yielded by more than one rater, inconsistencies might occur due to different criteria used by raters or differential application of the same criteria by the raters. Moreover, as Brown (2004) holds, the scoring process might also be affected by human error, subjectivity, and bias. The inconsistency of the scores yielded by different raters, according to him, may be due to factors such as undervaluing the scoring criteria, inattention, lack of experience, or preconceived biases.

Rater training is one of the recommended ways to overcome rater inconsistency and can serve as an effective tool to improve raters' agreement by creating a common understanding among them and introducing consistent scoring scales and criteria (Brown, 2004: Davis, 2016). It is considered as a way of mitigating the amount of variation in rater-related factors and a method for enhancing rater accuracy (Kang, Rubin, & Kermad, 2019). In training programs, raters are provided with the same sets of criteria based on which they judge individuals' performance and are given systematic training to be able to apply the criteria consistently (Brown,1995; Weigle, 1994). According to Lumley and McNamara (1995), following the introduction of the assessment criteria, some independent sessions might be allocated to rating a series of performances to calculate the inter-rater reliability to see how consistently the raters judge the performances. In relation to reliability, as Davis (2016) argues, the aim of rater training is to reduce the differences in scores from different raters. In this regard, he investigated the impact of experience and training on rater scoring patterns in the context of the TOEFL iBT Speaking Test. The participants were twenty English teachers who were inexperienced in scoring speaking within that specific research context. They holistically scored responses from

 ISSN 2476-5880

240 test-takers using the criteria listed on the iBT speaking scoring rubric. The raters completed one scoring session followed by one rater training session and three scoring sessions. Multifaceted Rasch analysis was used to analyze the severity and consistency of scores given by raters. The results indicated that training and experience had little impact on the variation in rater severity at the group level in general; however, it was found that some individual raters tended to score more severely than others, the differences of which were statistically significant. Also, it was found that inter-rater reliability and agreement moderately improved following the training sessions.

Kang, Rubin, and Kermad (2019) also investigated the effectiveness of rater training in diminishing interrater variance by examining how training might mitigate the amount of divergence between novice and experienced raters. The raters were eighty-two undergraduate or graduate students at a large Southeastern US university. Using the ETS holistic rating rubric for the TOEFL iBT speaking assessment, they rated speech samples from twenty-eight non-native speaker examinees responding to four TOEFL iBT speaking tasks once before receiving any training and once more following the training session. Analyzing the data to answer the research question concerning the degree to which a course of online rater training can neutralize the impact of potentially biasing rater background characteristics on speaking scores, they found that rater training is likely to "homogenize raters and improve the quality of their ratings by clarifying scoring criteria, reducing bias, and improving reliability" (p. 19). They concluded that training that provides raters with the scoring rubric and that exposes them to anchor performance samples can result in mitigating rater biases and making raters converge in their rating, at least for a period immediately after the training.

1.3. Ethics and fairness in language testing

Since Messick's (1989) introduction of 'consequential validity' in educational assessment, such issues as ethics in testing and assessment, fairness, and the reasonability of inferences about candidates based on test performance have become of great concern to testing professionals, teachers, and researchers. "Fairness is characterized by the absence of bias towards any identifiable group of test takers" (Amirian, Ghonsooly, & Amirian, 2020, p. 88). In language testing, fairness deals with the extent to which a score given to an individual language learner represents his/her real knowledge and skills as well as the extent to which a decision made for an individual language learner based on his/her performance is fair and bias-free (Brown, 2004). However, tests have consistently been used as a means of control and power throughout history (Spolsky, 1997), which implies that the issue of fairness has been almost neglected up to the present and the need for a sound, fair, and unbiased framework for language testing is now being strongly felt. In an attempt to partially integrate ethics in language testing, Shohamy (2001) suggests that raters make sure that their inclinations and feelings bias neither the assessment of individuals nor the decisions made based on their performance, which can be fulfilled by applying a set of appropriate performance criteria and standards that must be communicated to both teachers and students. Nevertheless, Hamp-Lyons (2001) argues that although this recommendation seems acceptable and reasonable, setting appropriate criteria

 ISSN 2476-5880

which apply appropriately to everyone with various learning preferences and different backgrounds might not be easily practical.

Despite all challenges involved in developing a coherent and reliable test fairness framework, Kunnan (2004, 2010) proposed a comprehensive framework with an ethics-inspired rationale. He claimed that the framework views fairness in terms of the whole system of a testing practice, not limited to the test itself. In an interview conducted by Jiang (2017) with Antony John Kunnan published in Language Assessment Quarterly, Kunnan asserted that fairness is "a fundamental aspect of a test" (p. 80) and one of the reasons people agree to take a test is because they suppose the test is going to be a fair one and that they do not want to take it if it is not fair. Kunnan's (2004, 2010) Test Fairness Framework (TFF), is presented in a linear list in Table 1.

Table 1
Kunnan's Test Fairness Framework adopted form Kunnan (2010)

Main quality	Main focus
1- Validity	
Content representativeness/coverage	Representativeness of items, tasks, topics
Construct or theory based validity	Representation of construct/underlying trait
Criterion-related validity	Score comparison with external criteria
Reliability	Stability, alternate form, internal consistency, inter-rater reliability
2- Absence of bias	Content or language bias
Content or language/dialect	Differential item functioning
Disparate impact	Criterion setting decisions
Standard setting	
3- Access	
Educational	Opportunity to learn
Financial	Affordable
Geographical	Location and distance
Personal	Accommodations
Equipment and conditions	Familiarity
4- Administration	
Physical setting	Physical conditions
Uniformity	Uniformity of administration
Security	Fraud, misrepresentation, cheating
5- Social consequences	
Washback	Effects on instruction
Remedies	Re-scoring, re-evaluation, legal remedies

ISSN 2476-5880

As indicated in Table 1, Kunnan's (2010) TFF considers validity as the first main component of the framework and defines it in terms of four types of evidence: Content representativeness/coverage evidence, construct or theory-based validity evidence, criterion-related validity evidence, and reliability. The second part of TFF has to do with the absence of bias. According to Kunnan (2010), the content or the language of the test might be offensive to specific groups of test-takers from different backgrounds and such aspects as age groups, gender type, race, native language, religion, or culture might be unfairly penalized by raters. Furthermore, Kunnan maintains that there might be items which favor a specific group of test-takers with a particular background, which requires a test to be examined for Differential Item Functioning (DIF). He also emphasizes that test developers and raters need to make sure the scoring criteria and standards they are using are unbiased enough to test true abilities and not construct-irrelevant factors. Based on the third part of the framework, test-takers need to have access to the test to be familiar with its content, different types of tasks, and cognitive demands of various kinds of test tasks. A test should be financially affordable and the test-taking location should be accessible to test-takers in terms of distance. It should also be appropriate for test-takers with physical or learning disabilities. Furthermore, test-takers need to be familiar with testing equipment, procedures, and conditions. The fourth part of the framework is concerned with the administration of the test which is defined in terms of three important considerations: Physical conditions (i.e., the appropriateness of the conditions for test administration), uniformity of test sites and equivalent forms and test manuals, and instructions, and test security (Kunnan, 2010). The final part of the framework concerns the effects of testing on instructional practices (e.g. teaching, materials, and test-taking strategies) or the remedies offered to test-takers to compensate for the detrimental consequences of a test. Kunnan's (2004, 2010) TFF is the theoretical backbone of the current study, the phases, and elements of which are described in the *Methods* section.

A lot of attention has recently been focused on the issue of test fairness since it is directly related to the validity of the test and test-takers (Amirian, Ghonsooly, & Amirian, 2020). In a recent study on fair student assessment, for instance, Murillo and Hidalgo (2020) carried out a phenomenographic study to investigate Spanish teachers' conceptions of what constituted a fair assessment. They collected data from thirty teachers at primary/secondary schools in Spain through a phenomenographic interview based on a self-report. They found two different conceptions about what a fair assessment was: one associating justice with equality and another connecting fair assessment with equity. That subgroup of teachers who linked justice with equality suggested four fundamental elements: transparency in information (the time, content, assessment tools, and criteria used by the teachers), the objectivity of tests (using highly structured methods that did not leave room for error), the eminently quantitative nature of assessment, and the training of students to enable them to perform well on the examinations they will have to take throughout their academic life. However, the other subgroup of teachers, who linked fair assessment with equity, asserted that assessments need to consider students' needs as well as their context (the type of environment where they grow up and develop), and similarly need to measure learning improvements using qualitative and procedural methods. They emphasized that fair assessment is student-centered and that, unlike homogeneous

assessments, the evaluation practice needs to be adapted to the characteristics and needs of students. Accordingly, as maintained by the teachers, fair assessments should measure students' efforts by considering their starting point and making continuous assessments to evaluate their learning progress, and quantitative instruments (e.g., exams) should be supplemented with qualitative information (e.g., through observation, teacher field diaries or portfolios). In conclusion, Murillo and Hidalgo (2020) asserted that if we follow an equality-based model, the result will be an education focused on the selection and ranking of students, which will feature social inequalities. However, following an equity-based model**,** as they argued, will lead to an education that will contribute to a more just society.

3. Statement of the problem and research questions

Concerning the assessment of oral language performance, several studies have so far addressed rater characteristics and rater bias (e.g., Lumley & McNamara, 1995; Wind & Peterson, 2018), variability in rater judgments (e.g., Barrett, 2001; Davis, 2016; Kang, Rubin, & Kermad, 2019), nonsystematic error in performance ratings (e.g., Henning, 1996; Kang, Rubin, & Kermad, 2019), rater training effects (e.g., Davis, 2016; Kang, Rubin, & Kermad, 2019), and rating scales in oral language assessment (Bland & Gareis, 2018; Chen, 2016; Jung Kim, 2006). Furthermore, recent research has focused on test fairness and social consequences of assessment from a theoretical point of view (e.g., Davies, 2008; Hamp-Lyons, 2001; McNamara, 2001; Murillo & Hidalgo, 2020), and some have developed conceptual theoretical frameworks for fairness in language testing (e.g., Kunnan, 2004, 2010). Nevertheless, according to Xi (2010), although previous fairness frameworks have been very useful in introducing general areas of potential research and practice, they "may not provide practical guidance on how to go about developing the relevant evidence to support fairness" (p. 148). Xi (2010) adds that the few empirical fairness studies have focused on only one of the various aspects of fairness at any one time. Moreover, to the best of our knowledge, no empirical studies in the field of oral assessment have addressed fairness from the test-takers' perspective. Therefore, besides investigating the effects of rater training on inter-rater reliability, this study aims to develop a valid and reliable questionnaire on fairness in oral language assessment based on Kunnan's (2004) TFF and investigates fairness from test-takers' point of view. The following research questions are thus raised.

1- *To what extent is rater training effective in promoting the inter-rater reliability of the oral language assessment of Iranian intermediate language learners?*

2. *Do rater training and the consideration of the examinees' expectations by the examiners have any effect on the test-takers' perception of being fairly evaluated?*

4. Method

4.1. Participants

The population of EFL learners from which the main participants were randomly selected included 8 intermediate EFL classrooms at various English Language Institutes of a western province of Iran (i.e., Kermanshah). After gaining their consent, 31 EFL learners (15 males and 16 females aging between 17 and 25) were selected to participate in the study. In addition,

 ISSN 2476-5880

74 students (33 males and 41 females aging between 16 and 28) from the same institutes were selected to be given an open-ended questionnaire to provide us with their expectations and impressions of fairness regarding oral language performance assessment. Also, to pilot the researcher-made questionnaire, 83 language learners (38 males and 45 females aging between 19 and 37) attending conversation courses or TOEFL/IELTS preparation courses at various private institutes completed the first version of the questionnaire.

The sample of raters included four experienced English teachers, three of whom were holding a master's degree in TEFL (Teaching English as a Foreign Language) and one of whom was a Ph. D. candidate in TEFL. They had taught English at various language institutes for more than 10 years and voluntarily agreed to participate in the study. In addition, 14 experienced EFL teachers (6 males and 8 females) participated in a focus-group interview session and provided the researchers with their impressions of fairness regarding oral language performance assessment.

4.2. Procedure

4.2.1. *Pre-training stage.* At this stage, the raters evaluated the oral language performance of 15 participants based on the IELTS speaking module interview. The context of the speaking test was precisely simulated to the real context of the IELTS. The questions were asked by all the four examiners, one part each. The examiners had been provided with scoring sheets with general criteria about pronunciation, vocabulary, and grammar accuracy as well as fluency and coherence to evaluate the participants' oral performance.

After the first oral performance examination, 74 EFL learners at the same institutes were given an open-ended questionnaire and were required to express their expectations and impressions of fairness in assessment with a special focus on oral performance tests. Moreover, a focus-group interview on the concept of fairness in oral performance assessment was conducted with 14 language teachers. Fitting students' and teachers' responses and comments into Kunnan's (2004) framework, a five-point Likert scale questionnaire (see Appendix), which was labeled Oral Language Assessment Fairness Questionnaire (OLAFQ), was developed to investigate fairness in oral-performance assessment from students' perspective. Four experts (university professors of TEFL) commented on the content and structure of the questionnaire. Based on their comments, a few items were deleted from and some items were added to the final battery of items. To ensure the validity and reliability of the questionnaire, it was administered to 83 language learners attending conversation or TOEFL/IELTS preparation courses at some private language institutes. Internal consistency reliability analysis and Principal Component Analysis (PCA) were run on the data obtained from the pilot study to ensure the reliability and construct validity of the questionnaire. Table 2 presents the result of Cronbach's alpha consistency reliability.

 ISSN 2476-5880

Table 2
Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.75	.75	53

As indicated in Table 2, the questionnaire enjoyed an acceptable level of reliability (α= .75). The results of Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity are presented in Table 3.

Table 3
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.602
Bartlett's Test of Sphericity	Approx. Chi-Square	4.83
	df	1378
	Sig.	.00

As shown in Table 3, the level of sampling adequacy was acceptable (KMO= .60) and Bartlett's Test of Sphericity was significant (Sig.= .00).

Finally, the examinees who participated in the first round of the speaking test were required to fill out the questionnaire. Their responses provided the researchers with their ideas on the extent to which they considered the oral performance test as a fair one.

4.2.2. Rater training stage. The training program was introduced to the raters in three separate one-hour sessions. Detailed information about performance standards and rating criteria was presented to the raters by one of the researchers. The IELTS band descriptors for the speaking test were fully discussed as well. Finally, to ensure the effectiveness of the training sessions, we showed the raters the video recordings of two candidates taking the speaking section of the IELTS and required them to rate their performance using the checklists. Then the ratings were analyzed and the inter-rater reliability was calculated, the results of which were almost satisfactory. The aim of the training program was to make raters aware of the criteria against which they could fairly evaluate EFL learners' oral performance and to help them use the same set of criteria to enhance rater consistency and promote inter-rater reliability.

4.2.3. Post-training stage. As mentioned above**,** before the administration of the second parallel speaking test, the OLAFQ was given to the EFL learners who had participated in the first speaking test and they were required to evaluate the extent to which they considered the examination process fair. Having analyzed their responses, the researchers informed the raters of EFL learners' impressions and expectations of a fair test of oral performance, and they were required to take the EFL learners' considerations into account in the upcoming second oral examination. Accordingly, the second group of participants (N= 16) were given the speaking subtest of IELTS but with modified content and procedure. Based on the examinees'

 ISSN 2476-5880

expectations and following Kunnan's (2004) TFF, students were informed of the performance standards and criteria against which their oral performance was to be assessed. The same examiners were required to rate the second group of participants' oral performance. However, to promote the reliability of the judgments, the test session was recorded in its entirety during the second interview and the raters were subsequently required to take a copy of the recording home and analyze it carefully before they score the final performances of the examinees. Then the inter-rater reliability was calculated once again to be compared with the previous index of interrater reliability to see whether there was any improvement as a result of the rater training program. Finally, OLAFQ was given to the second group of examinees and their responses were compared with those obtained from the first group to investigate possible changes in the examinees' impression and judgment of fairness.

5. Results

Table 4 shows the results of the ratings given by four raters to the performance of the first group of examinees on the speaking module of the IELTS.

Table 4
The results of the ratings at the pre-training stage

Examinees	Rater 1	Rater 2	Rater 3	Rater 4
1	5.5	6	5.25	4.25
2	7.25	7.25	6.25	8.5
3	8	7.75	8.25	7.75
4	8	6.25	6.75	6
5	6.5	6	7.25	7.75
6	6.75	6.25	6.75	5
7	7.5	6.75	6	7.5
8	6.75	6	6.5	7.75
9	7.5	6	6.25	7
10	6.75	7	6	7.75
11	4.75	4.25	3.75	4.25
12	6	5.25	3.75	5.5
13	6.5	5.5	4.25	6.75
14	6.5	6.75	6.5	7.75
15	6.25	6.75	6.75	6.5

Note: The scores range from 1 to 9.

To determine the consistency of the ratings at the first stage, an interrater reliability analysis using Intraclass Correlation Coefficient (ICC) was run in SPSS, the results of which are presented in Table 5.

 ISSN 2476-5880

Table 5
Alpha reliability statistics and Intraclass Correlation Coefficient for the first speaking test

Alpha reliability statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Raters
.86	.88	4

Intraclass Correlation Coefficient							
		95% Confidence Interval		F Test with True Value 0			
	Intraclass Correlation	Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.58	.33	.80	7.45	14	42	.00
Average Measures	.84	.66	.94	7.45	14	42	.00

As indicated in Tables 5, the ratings enjoyed a good level of inter-rater reliability (α= .86) and a good intraclass correlation coefficient level (ICC=.84), meaning that without any sort of training, the raters evaluated the examinees' performance fairly consistently.

The ratings of the second group of examinees at the post-training stage carried out by the same raters are presented in Table 6.

Table 6
The results of the ratings at the post-training stage

Examinees	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 4**
1	6.25	6	6	6
2	6.75	7	6.75	6.5
3	7	6.75	7	7
4	6	5	6.25	6
5	5.5	5.75	5.5	5.25
6	5.25	4.5	6	5.5
7	5	4.25	5.5	5.5
8	7.5	7	7.75	6.75
9	6.75	7	6.75	6.5
10	7.75	7.5	8	7.75
11	5	5.25	6.75	5.75
12	7.25	7.25	7.75	6.75
13	7	6.75	7.75	7
14	7	7	7.75	7.5
15	5.75	5	6.25	6
16	5.75	5	6.25	6

 ISSN 2476-5880

To answer the first research question (i.e. *To what extent is rater training effective in promoting the inter-rater reliability of the oral language assessment of Iranian intermediate language learners?*), another ICC interrater reliability analysis was run to determine consistency among raters in the second stage, the results of which are shown in Table 7.

Table 7
Alpha reliability statistics and Intraclass Correlation Coefficient for the second speaking test

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	Number of Raters
.95	.96	4

Intraclass Correlation Coefficient							
		95% Confidence Interval		F Test with True Value 0			
	Intraclass Correlation	Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.77	.54	.91	23.19	15	45	.00
Average Measures	.93	.82	.97	23.19	15	45	.00

As indicated in Tables 5, the ratings enjoyed an excellent level of inter-rater reliability (α= .95) and a strong intraclass correlation coefficient level (ICC=.93). Comparing the reliability statistics obtained for the ratings of the first group of examinees' performance with those obtained for those of the second group and based on the benchmark values proposed by Portney and Watkins (1993), it is concluded that the rater training program was quite effective in promoting inter-rater reliability coefficient from good (α= .86, ICC= .84) to excellent (α= .95, ICC= .93). This indicated that the training program could reduce rater variability and promote the inter-rater reliability to a satisfactory extent.

The PCA run on the original version of OLAFQ yielded thirteen factors; however, considering the interpretability of the factor solution and taking Kunnan's (2004, 2010) TFF as the basis of questionnaire development, we merged the related factors to obtain a well-organized, valid questionnaire assessing five factors and 15 subfactors. Table 8 indicates the descriptive statistics for all the items on the OLAFQ at both pre-training and post-training stages.

 ISSN 2476-5880

Table 8
Descriptive statistics for responses to OLAFQ at both pre-training and post-training stages

Factors	Items	Stage	Mean	SD	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)
1- **Validity**									
a) Content representativeness/coverage	1	Pre-training	3.47	.83	00.0	13.3	33.3	46.7	6.7
		Post-training	4.06	.68	00.0	00.0	18.8	56.2	25.0
	2	Pre-training	3.67	.81	00.0	6.7	33.3	46.7	13.3
		Post-training	3.75	.77	00.0	6.2	25.0	56.2	12.5
	3	Pre-training	3.40	.98	6.7	6.7	33.3	46.7	6.7
		Post-training	3.94	.85	00.0	6.2	18.8	50.0	25.0
b) Construct or theory-based validity evidence	4	Pre-training	2.87	.91	6.7	26.7	40.0	26.7	00.0
		Post-training	3.44	.96	00.0	18.8	31.2	37.5	12.5
	5	Pre-training	3.93	.70	00.0	00.0	26.7	66.7	6.7
		Post-training	4.25	.68	00.0	00.0	12.5	50.0	37.5
	6	Pre-training	3.80	.56	00.0	00.0	26.7	66.7	6.7
		Post-training	4.56	.62	00.0	00.0	6.2	31.2	62.5
	7	Pre-training	3.67	.81	00.0	6.7	33.3	46.7	13.3
		Post-training	3.88	.80	00.0	6.2	18.8	56.2	18.8
	8	Pre-training	3.73	.88	00.0	6.7	33.3	40.0	20.0
		Post-training	3.94	.92	00.0	6.2	25.0	37.5	31.2
	9	Pre-training	3.73	.79	00.0	6.7	26.7	53.3	13.3
		Post-training	4.00	.89	00.0	6.2	18.8	43.8	31.2
	10	Pre-training	3.53	.64	00.0	00.0	53.3	40.0	6.7
		Post-training	3.69	1.07	00.0	18.8	18.8	37.5	25.0
	11	Pre-training	2.87	.83	6.7	20.0	53.3	20.0	00.0
		Post-training	3.81	.75	00.0	6.2	18.8	62.5	12.5
	12	Pre-training	4.20	.56	00.0	00.0	6.7	66.7	26.7
		Post-training	4.56	.51	00.0	00.0	00.0	43.8	56.2
	13	Pre-training	4.20	.56	00.0	00.0	6.7	66.7	26.7
		Post-training	4.50	.63	00.0	00.0	6.2	37.5	56.2
	14	Pre-training	3.93	.70	00.0	00.0	26.7	53.3	20.0
		Post-training	4.50	.63	00.0	00.0	6.2	37.5	56.2
c) Criterion-related validity evidence	15	Pre-training	2.27	.79	13.3	53.3	26.7	6.7	00.0
		Post-training	2.19	.75	18.8	43.8	37.5	00.0	00.0
	16	Pre-training	4.13	.74	00.0	00.0	20.0	46.7	33.3
		Post-training	4.50	.51	00.0	00.0	00.0	50.0	50.0

 ISSN 2476-5880

d) Reliability	17	Pre-training	4.47	.64	00.0	00.0	6.7	40.0	53.3
		Post-training	4.69	.47	00.0	00.0	00.0	31.2	68.8
	18	Pre-training	2.00	.65	20.0	60.0	20.0	00.0	00.0
		Post-training	4.69	.60	00.0	00.0	6.2	18.8	75.0
	19	Pre-training	2.93	.70	00.0	26.7	53.3	20.0	00.0
		Post-training	4.19	.54	00.0	00.0	6.2	68.8	25.0
	20	Pre-training	3.00	.75	00.0	26.7	46.7	26.7	00.0
		Post-training	4.12	.61	00.0	00.0	12.5	62.5	25.0
	21	Pre-training	1.40	.50	60.0	40.0	00.0	00.0	00.0
		Post-training	1.88	.61	25.0	62.5	12.5	00.0	00.0
2- Absence of bias									
a) Offensive content or language	22	Pre-training	2.53	.91	13.3	33.3	40.0	13.3	00.0
		Post-training	1.87	.61	25.0	62.5	12.5	00.0	00.0
	23	Pre-training	1.87	.74	33.3	46.7	20.0	00.0	00.0
		Post-training	1.62	.61	43.8	50.0	6.2	00.0	00.0
	24	Pre-training	2.20	.86	20.0	46.7	26.7	6.7	00.0
		Post-training	1.81	.65	31.2	56.2	12.5	00.0	00.0
	25	Pre-training	1.73	.70	40.0	46.7	13.3	00.0	00.0
		Post-training	1.69	.70	43.8	43.8	12.5	00.0	00.0
b) Unfair penalization based on test-taker's background	26	Pre-training	1.80	.67	33.3	53.3	13.3	00.0	00.0
		Post-training	1.94	.68	25.0	56.2	18.8	00.0	00.0
	27	Pre-training	1.80	.67	33.3	53.3	13.3	00.0	00.0
		Post-training	1.62	.71	50.0	37.5	12.5	00.0	00.0
	28	Pre-training	3.93	.70	00.0	00.0	26.7	53.3	20.0
		Post-training	4.25	.77	00.0	00.0	18.8	37.5	43.8
3- Access									
a) Educational access	29	Pre-training	3.47	.74	00.0	6.7	46.7	40.0	6.7
		Post-training	4.19	.65	00.0	00.0	12.5	56.2	31.2
	30	Pre-training	3.67	.74	6.7	13.3	66.7	13.3	00.0
		Post-training	3.94	.77	00.0	6.2	12.5	62.5	18.8
b) Financial access	31	Pre-training	3.80	.86	00.0	6.7	26.7	46.7	20.0
		Post-training	3.88	.88	00.0	6.2	25.0	43.8	25.0
c) Geographical access	32	Pre-training	4.33	.61	00.0	00.0	6.7	53.3	40.0
		Post-training	4.25	.68	00.0	00.0	12.5	50.0	37.5
d) Conditions or equipment access	33	Pre-training	3.93	.70	00.0	00.0	26.7	53.3	20.0
		Post-training	3.94	.85	00.0	6.2	18.8	50.0	25.0

 ISSN 2476-5880

4- Administration									
a) Uniformity or consistency	34	Pre-training	3.80	.94	00.0	13.3	13.3	53.3	20.0
		Post-training	3.88	.80	00.0	6.2	18.8	56.2	18.8
	35	Pre-training	4.47	.64	00.0	00.0	6.7	40.0	53.3
		Post-training	4.44	.62	00.0	00.0	6.2	43.8	50.0
b) Physical conditions	36	Pre-training	3.73	.88	00.0	6.7	33.3	40.0	20.0
		Post-training	4.25	.68	00.0	00.0	12.5	50.0	37.5
	37	Pre-training	2.47	.91	13.3	40.0	33.3	13.3	00.0
		Post-training	3.69	.94	00.0	12.5	25.0	43.8	18.8
	38	Pre-training	4.27	.70	00.0	00.0	13.3	46.7	40.0
		Post-training	4.00	.73	00.0	00.0	25.0	50.0	25.0
5- Social consequences									
a) Washback	39	Pre-training	2.27	.79	13.3	53.3	26.7	6.7	00.0
		Post-training	2.06	.77	25.0	43.8	31.2	00.0	00.0
	40	Pre-training	2.00	.84	26.7	53.3	13.3	6.7	00.0
		Post-training	2.00	.81	25.0	56.2	12.5	6.2	00.0
b) Fairness of the decisions	41	Pre-training	2.40	.98	20.0	33.3	33.3	13.3	00.0
		Post-training	3.44	1.09	00.0	25.0	25.0	31.2	18.8
	42	Pre-training	2.20	.86	20.0	46.7	26.7	6.7	00.0
		Post-training	3.12	1.45	12.5	31.2	12.5	18.8	25.0
	43	Pre-training	3.20	1.08	6.7	20.0	26.7	40.0	6.7
		Post-training	4.06	1.06	12.5	12.5	31.2	43.8	00.0
c) Remedies	44	Pre-training	2.33	.81	13.3	46.7	33.3	6.7	00.0
		Post-training	4.25	.68	00.0	00.0	12.5	50.0	37.5
	45	Pre-training	2.00	.65	20.0	60.0	20.0	00.0	00.0
		Post-training	3.69	.79	00.0	6.2	31.2	50.0	12.5
	46	Pre-training	4.60	.50	00.0	00.0	00.0	40.0	60.0
		Post-training	4.81	.40	00.0	00.0	00.0	18.8	81.2

Since the data was found not to be normally distributed, a Mann-Whitney U test was utilized to answer the second research question (i.e. *Do rater training and the consideration of the examinees' expectations by the examiners have any effect on the test-takers' perception of being fairly evaluated?*), the results of which are displayed in Table 9.

 ISSN 2476-5880

Table 9
Mann-Whitney U test on the difference between examinees' perceptions of being fairly evaluated on their oral performance before and after the rater training program

Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)	Eta^2 (effect size)
730.50	1.81	-2.55	.01	.18

As indicated in Table 9, there was a significant difference between examinees' perceptions of being fairly evaluated on their oral performance before and following the rater training program (Asymp. Sig. = .011 < .05). Moreover, the difference was large enough in magnitude (Eta^2= .18) to allow the researchers to state confidently that the training program, along with the consideration of participants' expectations of a fair test, had been successful in enhancing the fairness of the examinees' oral language assessment.

Furthermore, to compare the first and the second groups of examinees' responses to individual items on the questionnaire, a Pearson chi-square analysis was conducted, the results of which are summarized in Table 10.

Table 10
Chi-Square Analysis on the OLAFQ items

	Percentage		**Chi-Square test**		
Numbers & Items of the Questionnaire	Group 1	Group 2	Chi-Square Value	df	Asmp. Sig. (2-sided)
6- Speaking fluently with only rare repetition or self-correction counted toward the final assessment.	6.7% Strongly Agree	62.5% Strongly Agree	10.80	2	.004
11- Sufficient knowledge of the target-language culture counted toward the final assessment.	20.0% Disagree	62.5% Disagree	10.02	4	.040
18- The interview session was audio-recorded in its entirety for further analysis and more precision.	00.0% Strongly Agree	75.0% Strongly Agree	27.99	4	.000
19- We were informed of the assessment criteria before the interview.	20.0% Agree	68.8% Agree	18.00	3	.000
20- We were informed of the performance standards before the interview.	26.7% Agree	62.5% Agree	13.33	3	.004
37- The test site and the interviewers' seating arrangements were stress-free.	13.3% Agree	43.8% Agree	9.86	4	.043
44- We were informed of our weaknesses and strengths after the interview.	46.7% Disagree	00.0% Disagree	21.72	4	.000

 ISSN 2476-5880

45- After were informed of our weaknesses, our teacher helped us eliminate them.	60.0% Disagree	6.2.0% Disagree	19.88	4	.001

As indicated in Table 10, a significant difference was found for items 6, 11, 18, 19, 20, 37, 44, and 45, suggesting a considerable difference between examinees' perceptions of being fairly evaluated on their oral performance before and after the rater the training program regarding these items. Items 6 and 11 are connected with "construct or theory-based validity evidence", items 18, 19, and 20 are associated with "reliability", item 37 is concerned with "test administration" and deals with the physical environment and the examiners' interpersonal behavior, and finally items 44 and 45 concern "the social consequences of testing". The consideration and refinement of the issues and conditions raised by these items at the post-training stage seemed to have considerably enhanced the examinees' feelings of being fairly evaluated.

6. Discussion

The results of the analyses indicated that without any training program raters had evaluated the examinees' performances fairly consistently. This consistency is probably justified by the purposeful selection of raters of similar background (e.g., expertise, university degree, age, native language, race, culture). As Barrett (2001) asserts, consistent rater performance depends on the selection process. In fact, the choice of raters is as important as the selection of tasks used in oral performance tests because different raters might differ in judging individuals' language performance, depending on their background (Davis, 2016) and the set of criteria they apply (Chalhoub-Deville, 1995; Kang, Rubin, Kermad, 2019). According to Lumley and McNamara (1995), rater characteristics (e.g., their degree of severity or harshness and their interpretations of the rating scales and scoring criteria) are likely to cause variability in test scores. In fact, raters might show some degree of severity toward particular groups of candidates, particular tasks, or particular rating occasions (Davis, 2016; Lumley & McNamara, 1995). The variability of raters' behavior needs to be compensated for in some way, the most prevalent method of which is rater training (Kang, Rubin, & Kermad, 2019). However, several other rater factors might be involved in contributing to variability among raters which could be avoided before the training program (Brown, 1995). Brown (1995) found that raters of particular backgrounds are likely to be more lenient than raters of other backgrounds and that they might be different in their perceptions of the importance of particular features of language and particular tasks. For instance, teacher raters in her study were found to be more concerned with and less tolerant of incorrect use of vocabulary items and linguistic structures than non-teacher raters. Thus, considering the attested effect of rater characteristics on the variability of the ratings, to homogenize raters and improve the quality of their ratings (Kang, Rubin, & Kermad, 2019), the raters in this study were deliberately selected from the same background.

The results of the analyses further indicated that the rater training program was effective in reducing rater inconsistency and promoting the inter-rater reliability coefficient from good to perfect. However, Brown (1995) found that, despite being exposed to training sessions, raters

differed slightly in the way they perceived and applied the assessment criteria probably because, based on their previous experience, they might have had inherent perceptions of what constituted an adequate knowledge of the language. She asserted that training raters and making the descriptors explicit to them were not enough to remove rater variability. Similarly, Davis's (2016) study indicated modest improvement in rater reliability and agreement following the training program. He claimed that rater training might create consistency in raters' own scoring but does not necessarily enhance inter-rater consistency. Nevertheless, the present study found a much stronger effect for rater training in that at the post-training stage, the raters evaluated examinees' performance considerably more consistently with a very small amount of variation than their earlier ratings.

Based on the results of this study, the refinement of "construct or theory-based validity evidence" seems to have enhanced the examinees' perceptions of being fairly evaluated. As mentioned earlier, the issue of test fairness has recently been the focus of attention because it is directly related to the validity of the test and test-takers (Amirian, Ghonsooly, & Amirian, 2020). Regarding the relationship between validity and fairness, Bachman (2000) asserts that ethics and validity issues in language testing are interrelated in such a way that the construct validity of interpretations cannot be investigated without considering the social values and consequences of the decisions made based on the test scores. Likewise, he maintains, the consideration of ethics without validity is impractical. He argues that to connect fairness with validation, we need a conceptual framework that is broad and flexible enough to encompass issues of validity, reliability, and the consequences and ethics of test use. Based on the same reasoning, the researchers in this study primarily built upon the principles of Kunnan's (2004) TFF, which is believed to be a comprehensive framework with an ethics-inspired rationale, and developed the OLAFQ and evaluated the examinees' perceptions of being fairly evaluated on their oral performance.

Another important quality contributing to test fairness was found to be the reliability of judgments and decisions, the consideration and refinement of which improved the examinees' perceptions of being fairly evaluated in this study. While assessing the EFL learners' speaking proficiency, it was quite a challenge for the raters to simultaneously focus on several aspects like interacting with the examinees, paying close attention to their oral performance, and scrutinizing their utterances and this challenge constituted a major threat to the reliability of the decisions made. To alleviate this problem, some examination boards hire two raters to do live ratings throughout the speaking test session, while others record the sessions for subsequent double-rating **(**Nakatsuhara, Inoue, & Taylor, 2021**).** Following these recommendations, to promote the reliability of the judgments, the test session was recorded in its entirety during the second interview and the raters were subsequently required to take a copy of the recording home and analyze it carefully before they allocated the final scores to examinees' performances. Another threat to the reliability of oral language assessment concerns the examinees' lack of knowledge about the criteria against which their performance is to be assessed. To eliminate this problem, the second group of examinees was informed of the criteria and expectations based on which their performance was supposed to be assessed. They were told that their performance would be evaluated based on criteria such as using a wide range of vocabulary

items, speaking naturally, coherently, and fluently, using communicative strategies to promote communication effectiveness or prevent breakdowns in communication, using a variety of grammatical structures, showing awareness of the target culture, etc.

Test administration, which deals with the physical environment as well as the examiners' interpersonal behavior, was also another condition that was considered and improved by the researchers in the second speaking test. The data obtained from the first group of examinees revealed that the majority of them felt uncomfortable with the testing environment, especially the seating arrangements of the examiners. Moreover, they found the idea of being simultaneously interviewed by four examiners rather threatening and anxiety-provoking. As emphasized by Donald (2016), individuals' attention and concentration are essential requirements of testing performance. Hence, in the second interview, three of the examiners were seated at one end of a larger table and one of them, who was asking the interview questions, was sitting in front of the examinees. In addition, to create a relatively stress-free environment, one of the examiners was nominated as the interview manager doing most of the interactions, and the other three examiners as raters were listening, taking notes, and filling out the checklists. The examinees' responses to the OLAFQ indicated that modifying the seating arrangement and interview procedure had remarkably reduced their stress and helped them feel more comfortable and perform more effectively.

"Consequential validity" was another factor the researchers considered to enhance the fairness of the assessment process. Subsequent to the second speaking test some remedies were offered to the participants by their teachers a few weeks after the examination as a sort of feedback in which their strengths and weaknesses were announced to them. Moreover, they were provided with some comments to help them eliminate their weaknesses. According to Gipps (1994, p. 137), "the requirement that students improve as a result of feedback can be seen as a consequential validity criterion for formative assessment". Gipps asserts that assessment is normally done for purposes such as supporting learning by giving teachers and learners detailed feedback. She considers feedback as a crucial feature of the teaching-learning process and adds that test results can be used by teachers to diagnose students' strengths and weaknesses, by students to determine their academic self-esteem, and by parents to monitor their children's progress. Accordingly, in this study, the researchers focused on the feedback provided by teachers to their learners to improve their subsequent performances. Hattie and Timperley (2007) conceptualize feedback as a consequence of performance and consider it as one of the most powerful influences on learning and achievement. To enrich the feedback provided to the examinees in this study, the teachers were required to incorporate students' common mistakes and problems into their material and deal with them continuously during the semester. In other words, in addition to providing feedback to the examinees on their performance, teachers set out to address the problems as a form of remedy during the course. Besides, students were exposed to some short lessons on pronunciation, communication strategies, and cultural issues during the course, which had a significant effect on their perception of test fairness.

In addition to the factors mentioned above, some other factors included in Kunnan's (2004) TFF were partly addressed to enhance the fairness of the decisions made based on the

test scores. For instance, to avoid using offensive or biased content or language, the topics were carefully selected so that they would be interesting and not religiously or culturally offensive. To avoid rater bias, particularly potential unfair penalization based on test-taker's background, the raters were purposefully selected to have the same native language as the examinees. Another considered factor was "access to the test". To ensure educational access, the topics were selected from among familiar topics discussed in class during their previous terms. To ensure financial and geographical access, the speaking module of IELTS was chosen to be utilized which was available at a reasonable price. Finally, factors related to "test administration" were taken into account. The interview was conducted in a quiet, well-lit room with the examiners sitting in a way that the examinees' feelings of discomfort and anxiety were alleviated. Finally, the interview site, test instructions, planning time, test length, and the raters were exactly the same for all the examinees to ensure the uniformity in test administration.

7. Conclusions and practical implications

In this study, the effect of rater training on inter-rater reliability was put under the magnifier and it was concluded that rater training was likely to reduce rater variability and, in turn, promote inter-rater reliability. This study also investigated test fairness from students' perspective and concluded that taking students' expectations of and impressions toward a fair test of oral language performance into account will improve their overall perceptions of being fairly evaluated.

The results of this study have implications for second language teachers. Regarding fairness in oral assessment, students expect to be informed of the criteria against which their performance is to be assessed as well as raters' expectations and the scoring standards before the interview. Furthermore, after being assessed, they expect to receive feedback on their performance, to be informed of their strengths and weakness, and to be given guidance on how they might eliminate their weaknesses and improve their future performances. Thus, to make their testing practices more meaningful and purposeful, teachers are advised to inform the students of their strengths and weaknesses during or following any assessment practice and help them improve their performance.

The results of this study also have implications for oral examiners and raters. To address ethical issues in testing and enhance fairness regarding test administration, oral examiners are recommended to guarantee testing conditions are the same for all individual test-takers. At the selection and planning level, they need to ensure uniformity in test administration, the interview site, the instructions, planning time, and test length. To promote rater consistency and invariability, they are advised to have some brainstorming or training sessions before the test to make sure they are applying the same scoring scale and criteria. This will contribute to obtaining a score that is closer to the examinee's true ability and is fair to him/her.

Finally, the results of this study have implications for oral test developers. One very critical point in oral assessment is the choice of topic. The interview questions are to include a wide variety of topics that are interesting, familiar, and bias-free to the target group and allow them to display their true speaking ability. They must also exercise caution not to include topics that might be offensive to particular groups of examinees. Another equally important issue is

content representativeness. Test developers are required to match the content of their tests with a comprehensive and operational definition of the construct of speaking ability to make sure the construct is not underrepresented.

One limitation of the present study was the number of examinees being evaluated on their oral performance. In this study, the performance of 15 examinees at the pre-training stage and that of 16 examinees at the post-training stage were evaluated. Further research is suggested to increase the number of examinees and raters to obtain more reliable results. If possible, future research is required to conduct the same study with IELTS candidates to evaluate their oral performance in the natural setting of the test and investigate their viewpoints on test fairness.

References

Amirian, M. R., Ghonsooli, B., & Amirian, K. (2020). Investigating fairness of reading comprehension section of INUEE: Learner's attitudes towards DIF sources. *International Journal of Language Testing*, *10*(2), 88-100.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1-42.

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

Bland L. M., & Gareis, C. R. (2018). Performance assessments: A review of definitions, quality characteristics, and outcomes associated with their use in k-12 schools. *Teacher Educators' Journal*, *11*, 52-69.

Brown, A. **(**1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing 12*(1), 1-15.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. White Plains, NY: Pearson Education.

Brown, J.D. & Bailey, K.M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34,* 21-42.

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies, 21*(2), 1-43.

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

Chen, G. (2016). Developing a model of analytic rating scales to assess college students' L2 Chinese oral performance. *International Journal of Language Testing*, *6*(2), 50-71.

Conn, C. A., Bohanb, K. J., Pieper, S. L., & Musumeci, M. (2020). Validity inquiry process: Practical guidance for examining performance assessments and building a validity argument. *Studies in Educational Evaluation*, 65, [100843]. https://doi.org/10.1016/j.stueduc.2020.100843.

Davies, A. (2008). Ethics, professionalism, rights and codes. In E. Shohamy & N. H. Hornberger, (Eds.), *Language testing and assessment: Encyclopedia of language and education*, 2nd edition, (Volume 7, pp. 429-443). New York: Springer.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, 33(1) 117-135.

Donald, E. K. (2016). Examinees' Perceptions of the Physical Aspects of the Testing Environment During the National Physical Therapy Examination (Doctoral Dissertation). University of South Florida, USA. Retrieved from https://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=7422&context=etd

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*(2), 125-144.

Gipps, C.V. (1994). *Beyond testing*. London: The Falmer Press.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Hamp-Lyons, L. (2001). Ethics, fairness, and developments in language testing. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honor of Allan Davies* (pp. 222-227). Cambridge: Cambridge University Press.

Henning, G. (1996). Accounting for nonsystematic error in performance ratings. *Language Testing, 13*(1), 53-61.

Jiang, J. (2017). Language Assessment: Critical issues—An interview with Antony John Kunnan. *Language Assessment Quarterly*, *14*(1), 75-88.

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing*, 36(4), 481-504.

Kazoki, Y. (2010). An alternative decision-making procedure for performance assessments: using the multifaceted Rash Model to generate cut estimates. *Language Assessment Quarterly*, 7, 75-95.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir, (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27-48). Cambridge: Cambridge University Press.

Kunnan, A. J. (2010). Statistical analysis for test fairness. *Revue Française de Linguistique Appliquée, 16,* 39-48.

Lumley, T. & McNamara, T.F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 55-71.

Luongo-Orlando, K. (2003). *Authentic assessment: Designing performance-based tasks*. Ontario, Canada: Pembroke Publishers.

McNamara, T. F. (2001). Language assessment as social practice: Challenges for research. *Language Testing, 18*(4), 333-349.

Messick, S. (1989). Validity. In R. L. Linn, (ed.), *Educational measurement*, 3rd edition. (pp. 13-103). New York: American Council on Education/Macmillan.

Murillo, F. J., & Hidalgo, N. (2020). Fair student assessment: A phenomenographic study on teachers' conceptions. *Studies in Educational Evaluation*, *65*, [100860]. https://doi.org/10.1016/j.stueduc.2020.100860

Nakatsuhara, F., Inoue, C., & Taylor, L. (2021). Comparing rating modes: Analysing live, audio, and video ratings of IELTS speaking test performances. *Language Assessment Quarterly*, *18*(2), 83-106.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*(4), 395-418.

Portney, L. G., & Watkins, M. P. (1993). *Foundations of clinical research applications to practice*. Connecticut: Appleton & Lange.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18*(4), 373-391.

Spolsky, B. (1997). The ethics of gatekeeping tests: What have we learned in a hundred years? *Language Testing, 14*(3), 242-247.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

Wesche, M. B. (1987). Second language performance testing: The Ontario Test of ESL as an example. *Language Testing, 4*(1), 28-47.

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2) 161-192.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147-170.

 ISSN 2476-5880

Appendix

Oral Language Assessment Fairness Questionnaire (OLAFQ)

Dear respondent,

The purpose of this survey is to investigate to what extent you perceive the oral test you have recently taken has been fair regarding a number of assessment-related factors. There will be no risks or negative consequences associated with participation in this research. The data obtained from this questionnaire will be treated as strictly confidential and no one, except the researchers, will have access to them. The questionnaire consists of 46 statements, and you are requested to rate each one on a 5-point Likert scale based on to what extent you believe the oral test you have recently taken has been fair.

Thank you very much for your sincere participation.

Strongly disagree= 1	Disagree= 2	No idea= 3	Agree= 4	Strongly agree= 5

1- Validity	**1**	**2**	**3**	**4**	**5**
a) Content representativeness/coverage					
1- The interview consisted of a wide variety of discussible topics.					
2-The several topics we were required to discuss during the interview gave us the chance to reveal our true speaking ability.					
3- The interview topics were familiar topics which had been discussed in class during the previous terms.					
b) Construct or theory-based validity evidence					
4- Sufficient knowledge and correct use of idiomatic vocabulary counted toward the final assessment.					
5- Using a wide vocabulary resource readily and flexibly counted toward the final assessment.					
6- Speaking fluently with only rare repetition or self-correction counted toward the final assessment.					
7- Speaking coherently with fully appropriate cohesive features and developing topics fully and appropriately counted toward the final assessment.					
8- Effective use of communication strategies to improve the efficiency of the communication counted toward the final assessment.					

 ISSN 2476-5880

9- Effective use of communication strategies to compensate for breakdowns in communication due to insufficient lexical or structural knowledge counted toward the final assessment.					
10- In addition to syntactic correctness of the utterances, their semantic meaningfulness also counted toward the final assessment.					
11- Sufficient knowledge of the target-language culture counted toward the final assessment.					
12- Using a full range of structures naturally and appropriately counted toward the final assessment.					
13- Producing error-free, grammatically correct sentences counted toward the final assessment.					
14- Using a full range of pronunciation features with precision and subtlety counted toward the final assessment.					
c) Criterion-related validity evidence					
15- The structure of the interview was similar to that of the language institute.					
16- The structure of the interview was similar to that of TOEFL and IELTS.					
d) Reliability					
17- The sufficient number of raters contributed to the reliability of the judgments.					
18- The interview session was audio-recorded in its entirety for further analysis and more precision.					
19- We were informed of the assessment criteria before the interview.					
20- We were informed of the performance standards before the interview.					
21- The interview was held more than once in order for the raters to make more reliable judgments.					
2- Absence of bias					
a) Offensive content or language					
22- The topics were not appropriate for our age group (i.e., young adults).					
23- Some of the interview topics were sexually offensive.					
24- Some of the topics were against our social and cultural values.					
25- Some of the topics were religiously biased.					
b) Unfair penalization based on test-taker's background					
26- The assessment of our performance has been biased by the gender differences between raters and the interviewees.					
27- The assessment of our performance has been biased by language or dialect differences between raters and the interviewees.					

 ISSN 2476-5880

3- Access					
a) Educational access					
28- We were familiar with the topics were required to discuss during the interview.					
29- Our teacher clarified the general structure of the interview in advance.					
b) Financial access					
30- The samples of the test (e.g., TOEFL & IELTS) were on the market and they were financially affordable.					
c) Geographical access					
31- The site where we could get the sample of the test was accessible in terms of distance.					
d) Conditions or equipment access					
32- We were familiar with test taking equipment such as computers.					
4- Administration					
a) Uniformity or consistency					
33- There was uniformity and consistency in test length.					
34- There was uniformity and consistency regarding the degree of difficulty of the tests.					
35- There was uniformity and consistency across test sites.					
b) Physical conditions					
36- The test was administered in a quiet and appropriate environment.					
37- The test site and the interviewers' seating arrangements were stress-free.					
38- The test was administered in a room with optimum light.					
5- Social consequences					
a) Washback					
39- Emphasizing the techniques and strategies needed for success on this interview, our teacher taught us in way during the term to prepare us.					
40- The materials and topics discussed during the term were completely in line with those needed for success on this test.					
b) Fairness of the decisions					
41- The result of the interview can be used to decide whether I have the ability promote to a higher level or not.					
42- The results of the interview to decide whether I'm qualified to study or to be employed in academic institutions.					
43- Overall, I consider my score fair regarding my true capabilities.					
c) Remedies					

ISSN 2476-5880

44- We were informed of our weaknesses and strengths after the interview.					
45- After were informed of our weaknesses, our teacher helped us eliminate them.					
46- The results of the interview were announced to us after a while.					