GRADIVA

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

# An Investigation of Rater's Assessment of Test Takers' Speaking Skill: The Effect of Gender and Delivery Platform

Nikoo Davarpanah[1], Mobina Masoomzade[2], Mobina Erfany[3], Mohammad Hossein Peravian[4], Houman Bijani[5]* & Mohammad Reza Oroji[6]

1,2.M.A., Department of English Language, Islamic Azad University, Zanjan Branch, Zanjan, Iran.
Email: [1]Nikoo.davarpanah@iauz.ac.ir, [2]Mobina.masoomzade@gmail.com
3.B.A., Department of English Language, Islamic Azad University,
Zanjan Branch, Zanjan, Iran. Email: 82.merfaninasab@gmail.com
4.PhD Candidate, Department of English Language, Islamic Azad University,
South Tehran Branch, Tehran, Iran. Email: Payam_peravian@yahoo.com
5,6.PhD, Department of English Language, Islamic Azad University, Zanjan Branch,
Zanjan, Iran. Email: [5]houman.bijani@gmail.com (*Corresponding Author),
[6]mohammadrezaoroji@yahoo.com

## Abstract

The present study was conducted in order to examine the effect of gender on the rating regarding oral proficiency and also the difference between video and audio performances of the EFL test takers. Three main instruments were used to achieve the mentioned goal. 60 EFL learners studying in Kish Institute in Zanjan were randomly selected to take a sample of TOEFL test including only listening and speaking skills. Oxford Placement Test was also given as a standard placement test in order to make sure of the participants homogeneity in terms of English proficiency. The participants were assigned into three group of levels based on their scores including elementary, intermediate and advanced. The speaking section of the test was rated by the researcher of the study along with one other colleague. The learners' performance was recorded in audio and video recording format. The speaking test included two tasks including a description and a narrative one. The raters were female and male as the factor of gender was considered as one of the main variables of the study. The data were analyzed through different statistics including two-way ANOVA and independent sample t-test and also intra class correlation was done to be sure of the consistency of the rating process. The study revealed that gender of the rater can have a significant effect on the scores they award to their learners while rating an oral proficiency on a speaking test. Also, it shed light on a significant difference between the audio and video performances on a speaking test as the delivery platform. The findings of the study are hoped to be beneficial in the field of language assessments in the context of Iran and the researcher hoped to broaden the related literature particularly regarding Iranian EFL learners and raters.

**Keywords:** *Assessment; Audio Performance; Rater, Gender; Oral Proficiency; Video Performance.*

## 1. INTRODUCTION

Foreign language assessment has been a field of challenges and controversies along the decades for teachers and students. Generally, foreign language classes are ruled by summative assessment practices aimed to measure learners' mastery of discrete language points and linguistic accuracy, rather than assessing students' communicative competence (Shaaban, 2005). However, although summative speaking assessment continues provoking reluctant attitudes in students, teachers may hardly approach this process differently, which may eventually lead learners either to succeed, fail or give up on the learning process.

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

Speaking is one of the most important skills to be developed and enhanced as means of effective communication. Speaking skill is regarded one of the most difficult aspects of language learning. Many language learners find it difficult to express themselves in spoken language. They are generally facing problems to use the foreign language to express their thoughts effectively. They stop talking because they face psychological obstacles or cannot find the suitable words and expressions. The modern world of media and mass communication requires good knowledge of spoken English.

In consequence, assessing the speaking skill is a complex process that requires special considerations for educators (Burns, 2012). For instance, teachers need to identify a suitable instrument or strategy that allows them to properly assess learners either "live" or through recorded performances (Ginther, 2012). Moreover, speaking assessment processes have to be closely related to teachers' instruction to help them make decisions considering students' linguistic abilities and course goals in order to select appropriate speaking tasks.

Regarding the matter of delivery platform in oral proficiency assessment, in the literature, contradictory views have been reported about the use of videos in listening tests. Shin (1998) found that when videos were used to assess listening, participants performed significantly better compared to an audio test group. Moreover, most (92%) test takers preferred listening assessment videos to audio (Progosh, 1996). On the other hand, Londe (2009) compared performances of test takers in two video formats (close-up of the lecturer's face and a full body view of the lecturer) against test takers in an audio-only format and found no significant differences between the three groups. The researcher claimed that the visual channel did not contribute to test-taker performance.

A brief look at the previous researches on oral proficiency assessment mainly focused on face- to- face assessment with taking gender factor into account, but to the best of the researcher's knowledge no investigation has been done on the effect of online (audio and video) platform. Therefore, according to the above-mentioned studies, there exists a gap in investigating the effect of test takers gender on oral proficiency assessment in online (audio and video) platform. Also, according to O' Loughlin (2002) to date, the role of gender in speaking tests has received limited attention in language testing research. It is possible in oral interviews, for instance, that both interviewing and rating may be highly gendered processes. In tests like the IELTS interview, where the interviewer also acts as the rater, this poses the question of whether a gender effect, if it exists, stems from the interview itself, the rating decision or a combination of both these events.

In order to achieve this goal 60 students to write a composition discussing advantages and disadvantages of online classes. Choosing the co-rater was based on two factors of academic education and teaching experience. In spite of a remarkable number of studies regarding the subject matter there are still some gaps which the present study was meant to focus on. Also, the reviewed studies reported different findings which was an area needed to be studies more precisely.

Although there have been claims on the relationship between gender and test performance (Aryadoust, 2016; O'Loughlin, 2002; O'Sullivan, 2000), little research has investigated the effect of test takers' gender differences on the raters' assessment of their oral language performance. Moreover, the impact of raters' gender differences on the consistency and severity measures of test takers' oral language assessment is unknown. Besides, measures of male and female raters' biases to the categories of the rating scale categories are still vague.

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

Also, the researcher has not been able to find any research finding which shows that the change and alteration of various elicitation technique in oral testing prompts may affect test takers' output and hence their scores. In this respect, although the results of some studies (e.g., Stansfield & Kenyon, 1992a) suggest different test performances on the audio oral proficiency interview (OPI) and simulated oral proficiency interview (SOPI) by test takers, it is not conclusive whether such differences are due to test format or other factors. That is, test takers themselves are not perfectly consistent; therefore, some variation might be due to the fact that certain test takers took one of the tests first and the other one next. In other words, there might be an interaction between test takers ability and the sequence of tests. Also, little conclusive research, so far, has investigated raters' degree of harshness when rating either methods of oral performance assessment.

The purpose of this study was aimed to investigate the effect of test takers gender and also delivery platform on oral proficiency assessment. Thus, this study was an attempt to accomplish the following aims: First, to find out the role of test takers and raters' gender when oral proficiency performances are tested. Second, to understand the differences or relations between audio and video performances of test takers. Second, to find other factors that may affect the oral proficiency assessment. Therefore the present study was aimed to answer the following questions;

1. Do raters' gender differences have an impact on the scores they award to test takers?
2. Is there any significant difference between audio and video oral performance in assessment?

## 2. LITERATURE REVIEW

O Loughlin (2002) reported his study in which data was collected for this study consisted of the audio-taped performances of 8 female and 8 male test-takers who undertook a practice IELTS interview on two different occasions, once with a female interviewer and once with a male interviewer. The interviews were transcribed and analyzed in relation to previously identified features of gendered language use, namely overlaps, interruptions and minimal responses. The scores later assigned by 4 raters (2 males and 2 females) to each of the 32 interviews were also examined in relation to the gender of both raters and test-takers using multi-faceted Rasch bias analyses. The results from both the discourse and test score analyses indicated that gender did not have a significant impact on the IELTS interview. These findings are interpreted in relation to more recent thinking about gender in language use.

Sandlund and Sundqvist (2016) examined empirical studies on L2 oral proficiency testing published between 2004 and 2014 with a particular focus on studies on discourse and social interaction in such tests. Interestingly, a majority of the studies examined did not include discourse data at all, which might be a reflection on authorship; that is, according to McNamara (2011) whether authors come from the field of measurement or applied linguistics. Searches also revealed that studies on OPIs were much more frequent than studies on paired or group tests. According to Ortega (2012) there was an increase of paired/group studies over the last few years of our set time frame, possibly mirroring the social turn within the broad field of second-language acquisition research.

Wu and Ortega (2013) describe a new Chinese Elicited Imitation Test (EIT) and reports on a study that investigated the degree to which it functions as a tool that can be used in second language acquisition research to gauge global second language (L2) oral proficiency. Eighty

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

L2 Chinese learners, sampled from two university curricular levels so as to represent high and low linguistic abilities and including both heritage and foreign language learners, participated in the study by completing the EIT as well as an oral narrative task and a background questionnaire. The results suggest that the new Chinese EIT can help measure overall oral linguistic proficiency in L2 Chinese for a variety of research purposes.

Henning (1983) Employed an initial sample of 143 adult Egyptian learners of English as a foreign language, the three oral testing methodologies of imitation, completion, and interview were compared for reliability and validity. Similarly, five components under each method, namely, raw score, fluency, pronunciation, grammar, and combined fluency-pronunciation-grammar ratings, were analyzed separately and in tandem. Multicomponent-multi method convergent and discriminant validities were determined. Stepwise multiple regression was computed using FSI-like interview scores as the dependent variable. And Rasch latent trait calibration and tests of fit validity were computed for imitation and completion tests.

Results indicated that the pronunciation component of the imitation method exhibited highest overall validity across all indexes. The FSI-like component of the interview method ranked second and the fluency component of the imitation method ranked third. Comparison of the three oral testing methods across all components for all empirical validity indexes showed (1) imitation, (2) interview, and (3) completion methods to rank in that respective order in terms of available composite validity indexes. Regression analysis showed the FSI-like interview to be primarily related to grammar skill from among 11 independent predictors examined.

Fortune et al., (2015) in their cross-sectional study used assessments developed by the Center for Applied Linguistics to examine the oral proficiency of 218 K–8 English-proficient students in 4 Spanish immersion programs. Following a comprehensive review of assessment results for English-proficient immersion learners, the article reports findings from statistical analyses. Ratings of student proficiency were significantly higher between Kindergarten and Grade 2 and between Grades 2 and 5; however, no significant differences were found between Grades 5 and 8, lending empirical support to the plateau effect identified in earlier immersion studies. Furthermore, positive moderate to strong correlations were found between teacher ratings and ratings assigned by trained assessment administrators. The article discusses implications for assessment tools and practices, immersion program design, and pedagogy.

Shohamy (1983) discussed the complexity of measuring oral proficiency in communicative situations. The difficulty is due to the large number of variables, linguistic and social, which interact with one another. It then reports on a study which examined the stability of the assessment of oral proficiency on the oral interview test. Students of Hebrew as a foreign language underwent four administrations of different versions of that test.

The administrations differed from one another by the occasion, the interviewer, the speech style, and the topic. Results from the analysis indicated that the different speech style and topic significantly affected students' scores on these tests while the occasion and the interviewer did not. The correlational analysis between pairs of tests pointed to low reliability and lack of stability of the tests, especially when two variables (i.e., occasion and tester) interacted. The results call for use of caution when decisions about individuals are made based on administration of communicative tests, for a need to identify sources of error in communicative tests, and for drawing stringent guidelines for the use of such tests.

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

Raters' scores may also be influenced by numerous intervening factors. Among these, personal factors such as gender, hunger, fatigue, illness, too bright or too dim light, room temperature or any disagreement with other raters have serious effect on test scores**.** 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). Among all rater effects, oral language teaching and rating experience are the variables which have attracted the most concentration. One of the most critical worrisome in raters' scoring is whether they have been adequately trained or have had enough expertise in assigning accurate scores**.**

A key issue which has frequently been shown to influence the assessment of learners' oral performance to a significant degree is the gender factor and gender-based perceptions and evaluations (Nakatsuhara, 2011; Porter, 1991). There have been a great number of research studies**.** On the relationship between language and gender (e.g., Aryadoust, 2016; O'Loughlin, 2002; O'Sulivan, 2000), which argued that the conversation styles of males and females are different. A majority of these studies claimed that females are more collaborative, cooperative and supportive than males when doing interactions. Some scholars, such as Sunderland (1995), even have gone far beyond claiming that men and women differ in terms of their communicative competence and assert that they have different norms of conversational interaction due to cultural, social and situational context variations. If such claims are true, then they will have important implications in the field of language assessment since they imply that oral language assessment is gender dependent.

Walt, Wet and Niesler (2008) investigated an attempt to use automatic speech recognition systems to obtain an objective score for oral proficiency and The process of test development and the subsequent digitalization of speech, trialing and evaluation will be discussed with specific reference to a course that leads up to a language endorsement required by teacher trainees in South Africa. Results show that the more specific rating instructions used in the second experiment improved intra-rater agreement, but made little difference to inter-rater agreement. In addition, the more specific rating criteria resulted in a better correlation between the human and the automatic scores for the repeating task, but had almost no impact in the reading task. Overall, the results indicate that, even for the narrow range of proficiency levels observed in the test population, the automatically derived ROS and accuracy scores give a fair indication of oral proficiency.

Betonio (2017) carried out the study in a Philippine state university to investigate if there is a significant difference between college students' English oral proficiency when they are grouped according to their current degree programs. Results show that there is a highly significant difference in the oral proficiency level of students in all areas, given by the significance value of 0.000 with 5% level of significance.

Rosane Silveira and Thaisy da Silva Martins (2020) investigated that how experienced raters use different types of scales to assess the development of oral proficiency in English as a second language (L2). These results may be partially due to the limited data provided to the raters and the small sample size. However, they still indicate that in formal language settings such as the one investigated here, where the communicative approach for language teaching prevails, special attention may need to be given to the teaching of grammar and pronunciation so that the development of these subcomponents can be enhanced.

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

## 3. METHODOLOGY

In order to investigate the research questions outlined in the first chapter of this thesis, the researcher employed a post-method research design in which a quantitative approach was used to investigate the raters' development over time with regard to rating L2 speaking performance (Cohen, Manion & Morrison, 2007). In addition, the type of sampling which was used in this study was "subjects of convenience", that is the subjects were selected based on certain reasons and they were not selected randomly (Dörnyei, 2009).

### *3.1. Participants*

To obtain reasonable answers to the research questions mentioned earlier, the following steps were taken: Test takers were selected randomly among those strata who were studying at the Kish Institute in Zanjan (In order to find out elementary, intermediate and advanced learners, Oxford Placement was administered to the participants through online platform. 60 adults Iranian EFL students, including 30 males and 30 females, ranging in age from 17 to 40 were willing to participate in the study. In other words, they were selected in a way that they represented three levels of language proficiency based on their class level placements and teachers' reports of their learning history; thus, their speaking ability levels were controlled while other student characteristics such as gender, age, native language, educational background and the number of years of probable residence in English speaking countries were not. The reason for choosing intermediate to advanced learners of English was that these students had already acquired the adequate knowledge regarding the required elements and standards of oral production. Among the many characteristics, the test takers' speaking ability was what the test intended to measure, and this was what the raters were supposed to focus on while scoring.

Two Iranian EFL teachers, including one male and one female ageing 35 and 42 were selected to participate in this study as the raters. These raters were experienced graduates of English language related fields of study, teaching in different schools, universities and language institutes. The raters in this study were selected based on availability at the time of the study and purposeful sampling (Dörnyei, 2009); that is, those who have already got the qualifications and of course were willing to participate took part in this study. The raters participating in this study were naturally both proficient but with a variety of levels of expertise; that is, the raters were different in terms of level of teaching, ranging from basic to advance. It should also be stated that both raters had high levels of English language proficiency although none was a native speaker of English language. In order to make the raters ready for assessing test takers, we were not going to train them at all, but for learning how to use the Speaking Rubric, we asked an experienced examiner as a benchmark to explain the procedure to them for one or two sessions.

### *3.2. Instruments*

1. Oxford Placement Test
2. The Speaking Test: The present study aimed to use the Community English Program (CEP) test to evaluate test takers' speaking ability under various language use situations.
3. The Scoring Rubric:

As one of the requirements of this study to evaluate the influence of using a scoring rubric on the validity and reliability of assessing test takers' oral performance, this study aimed to use an analytic rating scale.

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

### *3.3. Procedure*

The 60 EFL learners studying in Kish Institute in Zanjan were randomly selected to take a sample TOEFL test including only listening and speaking skills. Oxford Placement Test was given as a standard placement test in order to make sure of the participants homogeneity in terms of English proficiency. The participants were assigned into three group of levels based on their scores including elementary, intermediate and advanced. The speaking section of the test was rated by the researcher of the study along with one other colleague. The learners' performance was recorded in audio and video recording format. The speaking test included two tasks including a description and a narrative one. The raters were female and male as the factor of gender was considered as one of the main variables of the study.

## 4. RESULTS

In order to identify whether there exists a significant mean difference among the performance of the three groups of test takers in each tasks with respect to each of the basics of language analytical factors, a factorial MANOVA was used. Since there were 300 test takers participating in the study and 8 oral subcategory factors, 2400 data were obtained for data analysis. Table 4.4 displays the factorial MANOVA results of oral tasks and language analytical factors for the three groups of test takers.

**Table 1: Factorial MANOVA of Oral Tasks and Language Analytical Factors for the Three Groups of Test Takers**

| Source | Dependent Variable | Type III Sum of Squares | *df* | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| Corrected Model | Description | 37479.086[a] | 23 | 1629.525 | 2.027 | .003 |
| | Narration | 151507.785[b] | 23 | 6587.295 | 2.025 | .003 |
| | Summarizing | 342854.350[c] | 23 | 14906.711 | 2.028 | .003 |
| | Role Play | 610478.025[d] | 23 | 26542.523 | 2.026 | .003 |
| | Exposition | 955045.873[e] | 23 | 41523.734 | 2.026 | .003 |
| Intercept | Description | 644618.704 | 1 | 644618.704 | 801.855 | .000 |
| | Narration | 708159.615 | 1 | 708159.615 | 217.693 | .000 |
| | Summarizing | 782756.520 | 1 | 782756.520 | 106.477 | .000 |
| | Role Play | 854811.015 | 1 | 854811.015 | 65.252 | .000 |
| | Exposition | 936308.007 | 1 | 936308.007 | 45.678 | .000 |
| Test takers' levels | Description | 11422.226 | 2 | 1631.747 | 2.030 | .048 |
| | Narration | 47122.125 | 2 | 6731.732 | 2.069 | .044 |
| | Summarizing | 105327.943 | 2 | 15046.849 | 2.047 | .046 |
| | Role Play | 187217.865 | 2 | 26745.409 | 2.042 | .047 |
| | Exposition | 292771.947 | 2 | 41824.564 | 2.040 | .047 |
| Analytical factor | Description | 3257.108 | 7 | 1628.554 | 2.026 | .132 |
| | Narration | 13048.208 | 7 | 6524.104 | 2.006 | .135 |
| | Summarizing | 29690.801 | 7 | 14845.400 | 2.019 | .133 |
| | Role Play | 52907.520 | 7 | 26453.760 | 2.019 | .133 |
| | Exposition | 82784.241 | 7 | 41392.120 | 2.019 | .133 |
| levels * factors | Description | 22799.752 | 14 | 1628.554 | 2.026 | .013 |
| | Narration | 91337.452 | 14 | 6524.104 | 2.006 | .014 |
| | Summarizing | 207835.606 | 14 | 14845.400 | 2.019 | .013 |
| | Role Play | 370352.640 | 14 | 26453.760 | 2.019 | .013 |
| | Exposition | 579489.686 | 14 | 41392.120 | 2.019 | .013 |
| Error | Description | 1910089.210 | 2376 | 803.910 | | |
| | Narration | 7729164.600 | 2376 | 3253.015 | | |
| | Summarizing | 17466996.130 | 2376 | 7351.429 | | |

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

| | | | | | | |
|---|---|---|---|---|---|---|
| | Role Play | 31125810.960 | 2376 | 13100.089 | | |
| | Exposition | 48703188.120 | 2376 | 20497.975 | | |
| Total | Description | 2592187.000 | 2400 | | | |
| | Narration | 8588832.000 | 2400 | | | |
| | Summarizing | 18592607.000 | 2400 | | | |
| | Role Play | 32591100.000 | 2400 | | | |
| | Exposition | 50594542.000 | 2400 | | | |
| Corrected Total | Description | 1947568.296 | 2399 | | | |
| | Narration | 7880672.385 | 2399 | | | |
| | Summarizing | 17809850.480 | 2399 | | | |
| | Role Play | 31736288.985 | 2399 | | | |
| | Exposition | 49658233.993 | 2399 | | | |
| *a. R Squared = .019 (Adjusted R Squared = .010)* | | | | | | |
| *b. R Squared = .019 (Adjusted R Squared = .010)* | | | | | | |
| *c. R Squared = .019 (Adjusted R Squared = .010)* | | | | | | |
| *d. R Squared = .019 (Adjusted R Squared = .010)* | | | | | | |
| *e. R Squared = .019 (Adjusted R Squared = .010)* | | | | | | |

The outcome of the table demonstrates that there exists a significant difference among the performance of the three groups of test takers from each other (third row). This shows that the test takers, regardless of what subcategory factor is being considered, differed significantly from each other, $p<0.05$. However, considering the eight subcategory factors, there observed no significant difference among the test takers of in whole (fourth row).

This shows that the test takers, regardless of their proficiency different levels, did not differ from each other. Nevertheless, when considering both factors of test takers' various levels of proficiency and the eight different subcategory factors, there observed significant difference $p<0.05$ showing that the test takers of each level of proficiency performed differently from the other groups (fifth row) in each task with respect to the analytical factors of fluency, lexical complexity, structural complexity, and accuracy of oral language produced by the test takers.

RQ1: Is there any significant relationship between the raters' gender and the scores they award to test takers?

First the researcher tried to make sure of the existence of any relationship between the variables and then intended to examine the possible effect of the variable of gender on the scores the raters gave to their learner's performances.

**Table 2:** ***The Table of Pearson Correlation Test of the Two Variables***

| | | **Score** | **Gender** |
|---|---|---|---|
| Score | Pearson Correlation | 1 | .012 |
| | Sig. (2-tailed) | | .320 |
| | N | 30 | 30 |
| Gender | Pearson Correlation | .012 | 1 |
| | Sig. (2-tailed) | .320 | |
| | N | 30 | 30 |

Note. The relationship between the two variables of the study including the scores and the matter of the raters' gender was examined through running Pearson Correlation Test. According to the table above there was no significant relationship between the two mentioned variable. This indicates that raters' gender has no impact on the scores they awarded to the students'' oral performance.

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

**Table 3:** ***The Table of Intra-Class Correlation Coefficient of the Rater***

| | Intra-class Correlation | 95% confidence interval | | F Test with True Value 0 | | | |
|---|---|---|---|---|---|---|---|
| | | Lower Bound | Upper Bound | value | Df 1 | Df2 | sig |
| Single Measures | 133[a] | .10 | .363 | 2.078 | 16 | 96 | .015 |
| Average Measures | 519[c]. | .064 | .799 | 2.078 | 16 | 96 | .015 |

Note. According to the table above, there is an intra-rater reliability in the results of the raters scoring on the pretest and posttest in both control and experimental groups' scores. Therefore, the scores are proved to be rated in a consistent way by the rater of the present study.

Although the outcomes obtained above indicated that a number of both male and female raters were biased and inconsistent, it by no means indicate that their biases were due to test takers' gender differences or whether there was any significant relationship between raters' genders and their assessment of test takers' oral performances.

Therefore, in order to obtain convert the results obtained above in relation to the gender of raters with regard to the extent to which they showed interactional bias in scoring the test takers' performances, a second FACETS analysis was run. Table 4.106 displays an overall performance of male and female raters and their severity level scoring the test takers' of the same gender.

**Table 4:** ***Rater Gender Bias Measurement Report in Rating Male and Female Test-takers***

| Rater gender | Bias Measure (logits) | Z-score | SE | Infit Mn. Sq. |
|---|---|---|---|---|
| Male | 0.03 | 0.04 | 0.73 | 1.3 |
| Female | -0.23 | -0.31 | 0.74 | 1.3 |
| Mean | -0.10 | -0.13 | 0.73 | 1.3 |
| SD | 0.18 | 0.24 | 0.00 | 0.00 |
| Fixed (all same) chi-square = 2.27, *df* = 1, *p*>0.05 | | | | |

The outcome of the table revealed that both groups of raters, male and female, had relatively equal levels of severity to test takers. Since the obtained z-values were all within the acceptable range of ±2 (Wright & Linacre, 1994), it could be concluded that neither group of raters treated male or female test takers more harshly or more leniently with regard to gender similarities.

Also, in order to measure to what extent the raters of either gender group treated the treat takers with significant severity/leniency, their bias measures were analyzed. Bias measures which are in between the mean bias value ± half a logit value are considered as the acceptable severity/leniency value (McNamara, 1996; Wright & Linacre, 1994).

The mean bias measure was measured -10.0, thus those raters who displayed more than half a logit value above or below the mean logit value (between -0.60 and 0.40) would be considered as either too severe or too lenient.

Therefore, the above rater gender groups were both regarded to have an acceptable range of severity/leniency value showing that they did not show any significant severity or leniency to either male or female test takers. Moreover, in order to further ascertain that the little observed difference is not significant, a chi-square was used.

As shown in the table, the chi-square results indicated that there was no significant difference between the two groups of raters with regard their severity scoring test takers of the same gender (X2 (1, N=2) = 2.27, p>0.05).

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

**Table 4:** ***The table of Two-Way ANOVA***

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Corrected Model | 347.000[a] | 20 | 17.000 | 4.000 | .000 |
| Intercept | 10634.000 | 1 | 10634.000 | 2609.000 | .000 |
| proficiency | 267.000 | 13 | 20.000 | 5.058 | .120 |
| gender | .000 | 1 | .000 | .000 | .780 |
| proficiency * gender | 28.000 | 6 | 4.000 | 1.000 | .520 |
| Error | 268.000 | 66 | 4.076 | | |
| Total | 24683.000 | 87 | | | |
| Corrected Total | 616.000 | 86 | | | |
| a. R Squared = .564 (Adjusted R Squared = .431) | | | | | |

*Note*. The table above indicates the difference between the variables of the raters' gender and the scores of the learners having three proficiency level. The results revealed that we have no statistically significant interaction at the *p* level. We can see from the output above that there was no statistically significant difference in the effect of gender on the scores the raters gave to the learners having three proficiency levels. ($p < .0005$). Therefore, the first research null hypothesis was not rejected.

Is there any significant difference between audio and video oral performance in assessment?

**Table 5:** ***Independent samples T-test for speaking proficiency audio and video files***

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Sig. | T | Df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| | | | | | | | | | Lower | Upper |
| Grade | Equal variances assumed | 1.31 | .16 | -3.62 | 36 | .000 | 1.00 | .216 | -1.48 | -.517 |
| | Equal variances not assumed | | | -3.62 | 35.2 | .000 | 1.00 | .216 | -1.81 | -.568 |

The independent sample T-test procedure (table 4.5) offered two scoring results in two forms of audio and video reordered answering files. The significance index of the Levene statistic was .160 (greater than .05); it could be assumed that the both tests had equal variances. Based on Table 4.5, there was a significant difference (sig 2 tailed = .000) between the mean differences of the speaking proficiency test scores of the audio files and video file which were recorded to score the learners oral proficiency on doing two speaking tasks. Because the Sig (2-Tailed) value is less than 0.05. So, we can conclude that there is a statistically significant difference between two conditions ($p<0.05$). Thereby, the hypothesis that there is not statistically significant difference in terms of audio and video mode of answering files the score oral proficiency, was rejected.

A bias analysis was performed to analyze raters' behavior with respect to their severity, biasedness and consistency in scoring both test methods at the pre-training phase of the study (See Table 4.6). FACETS is capable of calculating raters' biases in various testing contexts – in particular here test methods (audio and video tests) by comparing the expected and observed values in a set of data and then reporting the outcome in a form of residuals. Later on, through converting residuals into –scores, the bias value is obtained. This z-score shows any significant

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

difference from what was expected from that particular rater allowing for routine and acceptable score variation. Finally, a z-score in between ±2 is regarded as a rater's normal scoring behavior thus acceptable range of biasedness. Column one (Oral test method) displays the oral test methods used in the study, i.e., audio and video test method. Column two (Observed average score) displays the average observed scores given by the raters to test takers' oral performance on each test method.

Column three (Fair average) demonstrates the extent to which the mean ratings of raters on each test method differ. For instance, here, the mean rating of the audio test method was 22.12 and its fair average was 22.90. Similarly, the mean rating of the video test method was 18.74 and its fair average was 19.82. These data show that the two test methods were 0.78 raw-score points apart when comparing their mean ratings and 1.08 raw scores apart when comparing their fair averages. According to Winke, Gass and Myford (2012) both values demonstrate severity spread; however, the difference is that fair average is a better estimate when not all raters scored all the tasks. Wolfe and Dobria (2008, cited in Winke, Gass & Myford, 2012) further reiterated that when fair average is greater than 1 point, then this shows a significant high difference between the severest and the most lenient raters in the use of scoring scale. Column four (Obs-Exp score in logits) displays the total observed score for all the 100 test takers participating at the pre-training phase of the study on each test method minus the total expected score for the test takers on the same test method. Since there were 5 tasks in the study for each test method and for each task, the allowed score range was between 1 and 7, there would be the possibility of scoring each test taker a score of 5 to 35. Therefore, As an example, if a test taker whose expected score is 26 receives 21 from a rater, then the difference will be (21-26 = -5). Then, indicated value on the table is the obtained score in logits.

Column five (Bias logit) demonstrates the bias value, representing raters' severity/leniency (in each test method) in the performance assessment of test takers of that test method. Positive values represent severity, whereas negative ones represent leniency. Here, the outcome shows that the raters in the audio test method were rather lenient towards the test takers with the leniency of (-0.41 logits). However, rather in the video test, on the other hand, were severe with the severity of (0.17 logits). The mean bias value (in logits) measured -0.09, thus the raters in either test method who displayed more than half a logit value above or below the mean logit value (between -0.59 and 0.41) would be considered as either too severe or too lenient (McNaramar, 1996; Wright & Linacre, 1994). In this respect, no significant severity/leniency was observed on the ratings of the either audio or video test method, in other words, and the obtained severity estimate was within the acceptable range. Column six (SE) displays the standard error of bias estimation. The small amount of SE provided evidence for the high precision of measurement.

Columns seven and nine (Infit and Outfit mean square) display the fit statistics which show to what extent the data fit the Rasch model, or the difference between the observed scores and the expected ones. An observed score is the one given by a rater to a test taker on one criterion for a task, and an expected score is the one predicted by the model considering the facets involved (Wright & Linacre, 1994). In other words, fit statistics simply is used to determine within-rater consistency (Intra-rater consistency) which indicates the extent to which each rater ranks the test takers consistent with his/her true ability. Fit statistics is categorized into two subparts entitled infit and outfit statistics and most researchers employ them because they are said to be less sensitive to sample size and that they are commonly weighted on the information provided by the responses. Infnit is the weighted mean square statistic which is

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

weighted towards expected responses and thus sensitive to unexpected responses near the point where the decision is made. In other words, it is the average difference between actual scores and the estimated scores provided by the analysis. Outfit is the same as above but it is unweighted and is more sensitive to sample size, outliers and extreme ratings (Bonk & Ockey, 2003). Fit statistics has the expected value of 1 and a range of zero to infinity; however, there is no straightforward rule, absolute or universally definite range for interpreting fit statistics value or for setting the upper and lower limits; therefore, the acceptability of fit is done on a judgmental basis not solely on a statistical one. The acceptable range of fit statistics, although various among statisticians, according to Wright and Linacre, (1994), is within 0.6 to 1.4 logit values. Therefore, in order to investigate the fit statistics value. The raters (of each test method) who are placed below this range are overfit or too consistent, and those above this range are underfit (misfit) or too inconsistent. The infit mean square for the audio test method measured 1.2 and for the video method 1.3. This finding demonstrates that the ratings of both test methods, according to Wright and Linacre (1994), are within the acceptable fit statistics range showing relative consistency before training, however, for the video test method, through considering the outfit mean square value, they were spotted on the borderline of consistency.

Also, columns eight and ten (Z-scores) which are sometimes called standardized infit statistics display test method rater bias estimate at this phase of the study. Bias is the difference between expected and observed ratings of the obtained data which is then divided by its standard error to achieve then z-score (Stahl & Lunz, 1992). The most preferable amount of z value is 0 which indicates that the data match the expected model, thus there exists no bias on the side of raters. The z-scores are also plotted into a graph showing raters' biasedness map in each test method at each phase of the study. The maps are provided at the end of delayed post-training data analysis of this section. According to McNamara (1996), z values between ±2 are considered as the acceptable range of biasedness, thus any values above or below the given score are considered to be either to positively biased or too negatively biased. Accordingly, the raters in both test methods were considered as having nonsignificant biasedness but to opposite audioions. i.e., the raters of the audio audio method had the tendency towards leniency ($Z_{Audio}$= -0.81) while for the raters of the video test, the tendency was towards severity ($Z_{Video}$= 0.66). Although the ratings of both test methods were within the acceptable range of biasedness, the result indicates that the amount of biasedness for the ratings of the audio test method at the pre-training phase of the study was more than that of the video test method.

However, the logit severity estimates do not themselves tell us whether the differences in severity/leniency estimates are meaningful or not; consequently, FACETS also provides us with several indications of the reliability of differences among the elements of each facet. The most helpful ones to study are Separation index, Reliability and Fixed chi-square which can be found at the bottom of the table. The separation index is the measure of the spread of the estimates related to their precision. In other words, it is the ratio of the corrected standard deviation (usually written in short Adj. SD.) of element measures to the root mean square estimation error (RMSE) which shows the number of statistically distinct levels of severity among the raters. In case the raters were equally severe, the standard deviation of the rater severity estimates should be equal to or smaller than the mean estimation error of the entire data set which results in a separation index of 1.00 or even less (if there is a total agreement among raters, the separation index should be 0.00). In the case of this phase of the study, the separation index of 2.73 for the audio test method and 2.44 for the video test indicated that the variance among the raters, of each test method, was more than the error of estimates. This shows that the raters of each of the test methods were not equally severe.

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

The reliability demonstrates to what extent or how well the analysis distinguishes among the facet elements with respect to various levels of severity/leniency. It is exactly the same as Cronbach alpha in classical true score (CTS). It is noteworthy to indicate that, for this analysis, a low reliability (for all facets except test takers) is desirable, because in an ideal situation various raters would have equal amount of severity thus the analysis would not be able to distinguish the severe raters from lenient ones. However, the high amount of reliability in video test method, indicated that the analysis could reliably separate the raters of each test method into different levels of severity. Fixed chi-square tests the null hypothesis to check whether all elements of the facet are equal or not. The fixed chi-square value for all the 20 raters rating the test takers' oral performance of each test method was measured. The chi-square value indicated that there was significant difference in raters' level of severity ($X2$ (1, N=2) = 87.64, $p< 0.00$). Here, the high value of chi-square indicated that the ratings of the two test methods did not share the same on a parameter (e.g., severity). Consequently, the outcome suggested that the raters of either test method were not at the same level of severity.

As it was already indicated above, the raters' separation indices which were measured 2.73 and 2.44 for the audio and video test methods respectively indicated that there were almost three statistically distinct levels of severity. Statistically distinct levels are defined as those separation indices that are three standard errors apart, centered on the mean of the sample (Wright & Masters, 2002, cited in Davis, 2015). The reliability of this rater separation indices were 0.91 and 0.94 for the audio and video test methods respectively showing that the raters were reliably separated with respect to their level of severity and the analysis was reliable. As explained by Wink, Gass, and Myford (2012) separation reliability indices close to zero show that raters did not differ significantly in terms of their levels of severity and that they had rather similar levels of severity; whereas the separation reliability indices close to 1.0 demonstrate that the raters were very reliably separated with respect to their severity levels. Here, the rater separation reliability of 0.91 and 0.94 for audio and video test methods represents that the raters differed with regard to their severity variation in scoring the examinees oral performance. Column eleven (Point biserial correlation) displays the correlation coefficient between each rater and the rest of the raters participating in this study in either of the test methods. Here, the correlation coefficient for the audio test method was measured 0.26 (less than typical according to Cohen's table of effect size) and for the video method 0.38 (typical according to Cohen's table of effect size).

**Table 6:** ***Audio and Video Test Methods Measurement Report (Pre-training)***

| Oral test method | Observed average score | Fair average | Obs-Exp score (logit) | Bias Logit | SE | Fit statistics | | | | Correlation Point biserial |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Infit MnSq. | Z | Outfit MnSq. | Z | |
| Audio | 22.12 | 22.90 | 0.75 | -0.41 | 0.05 | 1.2 | -0.81 | 1.1 | -0.77 | 0.26 |
| Video | 18.74 | 19.82 | -0.31 | 0.23 | 0.04 | 1.3 | 0.66 | 1.4 | 0.43 | 0.38 |
| Mean | 20.43 | 21.36 | 0.22 | -0.09 | 0.04 | 1.25 | -0.07 | 1.25 | -0.17 | 0.32 |
| SD | 2.39 | 2.17 | 0.74 | 0.45 | 0.00 | 0.07 | 1.03 | 0.21 | 0.84 | 0.08 |

## 5. DISCUSSION

This study aimed to examine the impact of raters' gender on their scoring of learners' oral proficiency and whether differences exist between scores given to audio and video-recorded performances. The results indicated no significant gender-based scoring differences. This

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

aligns with studies by O'Loughlin (2002), Lumley and Sullivan (2005), and O'Sullivan (2002), which also found no gender-based bias in oral performance ratings. However, other studies (e.g., Aryadoust, 2016; O'Sullivan, 2000) reported small but significant gender effects on test scores. Some research (Nakatsuhara, 2011; Porter & Shen, 1991; Buckingham, 1997) suggested that raters tend to favor test takers of the same gender. Contradictory findings may be attributed to differences in statistical analysis methods or cultural factors, as some prior studies focused on Japanese, Arab, and Indian contexts.The study also investigated whether raters' expertise and gender influenced their bias toward male or female test takers. No evidence of bias was found, contradicting research (e.g., Lim, 2011; Winke, Gass, & Myford, 2012) that suggested experienced and inexperienced raters judge performances differently.

Furthermore, neither male nor female raters demonstrated significant leniency or harshness toward test takers of the same or opposite gender. Since past studies have yielded mixed results, further research is recommended. The study also examined differences in scoring between test takers' audio and video-recorded performances. The results were mixed, with some studies (e.g., Progosh, 1996; Shin, 1998; Sueyoshi & Hardison, 2005) suggesting video assessments were more effective, while others (e.g., Londe, 2009; Gruba, 1993) found no significant differences. The present study found that audio and video performances were similarly rated, but video presentations did not necessarily enhance comprehension. In terms of test design, oral assessments should consider validity, feasibility, and fairness (Kenyon & Tschirner, 2000).

The study confirmed that well-designed oral tests are stable, reliable, and valid if they align with real-life language use. However, test takers found semi-audio oral tests more stressful than audio-only ones, a finding consistent with prior research (Jeong & Hashizume, 2011). Audio oral tests were more appropriate for lower-level test takers, while video oral tests were better suited for higher-ability learners. Linguistic differences were noted between audio and video oral tests. Video test takers made more pronoun errors, while audio test takers had more tense and verb structure errors, supporting Shohamy's (1994) findings. Video test takers also used self-correction more frequently, possibly due to a heightened focus on linguistic accuracy. Audio test takers, in contrast, paraphrased more, possibly because of limited direct interaction with an examiner. This aligns with Tarone's (1983) theory of Interlanguage Continuum, suggesting video tests elicit more structured speech, while audio tests resemble natural communication. In sum up, the study found no gender bias in scoring and no significant difference in assessments of audio versus video performances. However, given the diversity in previous research findings, further studies are recommended to clarify these issues.

## 6. CONCLUSION

The study revealed that gender of the rater can have a significant effect on the scores they award to their learners while rating an oral proficiency on a speaking test. Also, it shed light on a significant difference between the audio and video performances on a speaking test as the delivery platform. The findings of the study are hoped to be beneficial in the field of language assessments in the context of Iran and the researcher hoped to broaden the related literature particularly regarding Iranian EFL learners and raters. Unlike most previous studies which have suggested the direct and most significant cause of test takers' score variability due to task difficulty level, the finding of the study provided evidence on the higher influence of test takers' own ability in their oral score variance. The findings of this study also suggested that gender, either on account of the raters or the test takers sides, does not have any significant impact, on

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

the one hand, on the performance ability of the test takers and on the other hand, the biasedness of raters' scoring patterns.

## 7. SUGGESTIONS FOR FURTHER STUDIES

The future studies can focus on the matter of the learners' gender. Also, they can be carried out with a larger sample of the learners and the raters. Also, questionnaires can be useful in order to elicit the raters and learners' attitudes towards the studies issues such as delivery platform and the advantages and disadvantages through interviews or questionnaires.

## References

1) Aryadoust, V. (2016). Gender and academic major bias in peer assessment of oral presentations. *Language Assessment Quarterly*, *13*(1), 1-24.

2) Betonio, H. R. Assessment of Students' English Oral Proficiency Based on Degree Programs: Implications for Admission Examinations.

3) Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89-110.

4) Burns, A. (2012). A holistic approach to teaching speaking in the language classroom. In Symposium (pp. 165-178)

5) Cohen, L., Manion, L., & Morrison, K. (2017). Action research. In *Research methods in education* (pp. 440-456). Routledge.

6) Dörnyei, Z. (2009). The L2 motivational self-system. *Motivation, language identity and the L2 self*, *36*(3), 9-11.

7) Fortune, T. W., & Tedick, D. J. (2015). Oral proficiency assessment of English-proficient k–8 Spanish immersion students. *The Modern Language Journal*, *99*(4), 637-655.

8) Ginther, A. (2012). Assessment of speaking. *The Encyclopedia of applied linguistics*, 1-8.

9) Gruba, P. (1993). A comparison study of audio and video in language testing. *JALT journal*, *15*(1), 85-88.

10) Henning, G. (1983). Oral proficiency testing: Comparative validities of interview, imitation, and completion methods. *Language learning*, *33*(3), 315-332.

11) Kenyon, D. M., & Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *The Modern Language Journal*, *84*(1), 85-101.

12) Jeong, H., Hashizume, H., Sugiura, M., Sassa, Y., Yokoyama, S., Shiozaki, S., & Kawashima, R. (2011). Testing second language oral proficiency in direct and semidirect settings: A social-cognitive neuroscience perspective. *Language learning*, *61*(3), 675-699.

13) Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language testing*, *28*(4), 543-560.

Volume 64 | Issue 03 | March 2025
DOI: 10.5281/zenodo.15017258

14) Londe, Z. C. (2009). The effects of video media in English as a second language listening comprehension tests. *Issues in Applied Linguistics*, *17*(1).

15) Lunz, M. E., & Stahl, J. A. (1992). New ways of thinking about reliability. *Professions Education Researcher Quarterly*, *13*(4), 16-18.

16) McNamara, T. (2011). Measuring deficit. In *Discourses of deficit* (pp. 311-326). London: Palgrave Macmillan UK.

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

18) O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language testing*, *19*(2), 169-192.

19) Ortega, L. (2012). Epilogue: Exploring L2 writing–SLA interfaces. *Journal of Second Language Writing*, *21*(4), 404-415.

20) O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System*, *28*(3), 373-386.

21) Porter, D. (1991). Affective factors in the assessment of oral interaction: Gender and status.

22) Progosh, D. (1996). Using video for listening assessment: Opinions of test-takers. *TESL Canada Journal*, 34-44.

23) Sandlund, E., Sundqvist, P., & Nyroos, L. (2016). Testing L2 talk: A review of empirical studies on second-language oral proficiency testing. *Language and Linguistics Compass*, *10*(1), 14-29.

24) Shaaban, K. (2005). A proposed framework for incorporating moral education into the ESL/EFL classroom. Language, Culture and Curriculum, 18(2), 201-217**.**

25) Shohamy, E. (1983). The stability of oral proficiency assessment on the oral interview testing procedures. *Language Learning*, *33*(4), 527-540.

26) Silveira, R., & Martins, T. D. S. (2020). Assessing second language oral proficiency development with holistic and analytic scales. *Ilha do Desterro*, *73*, 227-250.

27) Stansfield, C. W., & Kenyon, D. M. (1992). The development and validation of a simulated oral proficiency interview. *The Modern Language Journal*, *76*(2), 129-141.

28) Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language learning*, *55*(4), 661-699.

29) Tarone, E. (1983). On the variability of interlanguage systems. Applied linguistics, 4(2), 142-164**.**

30) Walt, C., de Wet, F., & Niesler, T. (2008). Oral proficiency assessment: the use of automatic speech recognition systems. *Southern African Linguistics and Applied Language Studies*, *26*(1), 135-146.

31) Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language testing*, *30*(2), 231-252.

32) Wright, B. D., & Linacre, J. M. (1994). The Rasch model as a foundation for the Lexile Framework. *Unpublished manuscript*.