Journal of Teaching Language Skills (JTLS)
36(3), Fall 2017, pp. 1-32- ISSN: 2008-8191
DOI: 10.22099/jtls.2017.24841.2223

The Impact of Raters' and Test Takers' Gender on Oral Proficiency Assessment: A Case of Multifaceted Rasch Analysis

Houman Bijani
PhD Candidate
Islamic Azad University, Science and Research Branch
houman.bijani@gmail.com

Mona Khabiri*
Associate Professor
Islamic Azad University Central Tehran Branch
monakhabiri@yahoo.com

Abstract

The application of Multifaceted Rasch Measurement (MFRM) in rating test takers' oral language proficiency has been investigated in some previous studies (e.g., Winke, Gass, & Myford, 2012). However, little research so far has ever documented the effect of test takers' genders on their oral performances and few studies have investigated the relationship between the impact of raters' gender on the awarded scores to male and female test takers. Thus, this study aimed to address the above-mentioned issue. Twenty English as a Foreign Language (EFL) teachers rated the oral performances of 300 test takers. The outcomes demonstrated that test takers' gender differences did not have any significant role in their performance differences when they were rated by the raters of the same or opposite gender. The findings also reiterated that raters of different genders did not demonstrate bias in rating test takers of the opposite or same gender. Moreover, no significant difference was observed regarding male and female raters' biases towards the rating scale categories. The outcomes of the study showed that both male and female raters assign fairly similar scores to test takers. This suggests no evidence based on which either male or female raters must be excluded from the rating process. The findings imply that there is no need to worry about the impact of gender for a more valid and reliable assessment.

Keywords: bias, multifaceted Rasch measurement (MFRM), rating scale, severity

Received: 27/06/2017 Accepted: 23/12/2017

* Corresponding author

Speaking skills are major parts of the curriculum in language teaching and this makes them an important subject of assessment accordingly (Baleghizadeh & Gordani, 2012). As pointed out by Bachman (2004), one of the areas of difficulty in language testing has always been the measurement of oral proficiency. Assessing speaking, according to Fulcher, Davidson and Kamp (2011), is challenging because there are so many factors that influence our impression of how well someone can speak a language, and it requires students to produce complex answers integrating language skills. A major difficulty is that, we expect that oral test scores, which are typically conducted by one or more human raters, to be accurate and appropriate for our purposes. This is a big expectation, and in different contexts teachers and testers have tried to achieve all these through a range of different procedures (Luoma, 2004). Consequently, these two reasons, the test method and the raters, make the assessment of speaking a real challenge.

One of the most prominent concerns, no matter how carefully the test is constructed, is the issue of raters. It has long been understood that test score variability is mainly related to rater factors (Barkaoui, 2011). McNamara and Lumley (1997) stated that, "the study found the element of chance in these public examinations to be such that only a fraction of the successful candidates can be regarded as safe, if a different set of equally competent judges had happened to be appointed" (p. 55). The reliability of a rating scale is also dependent on the raters who operate it (Fulcher, Davidson, & Kamp, 2011). In other words, a major concern with performance assessment is that the tasks require subjective assessment by raters. So, raters, as an additional source of measurement error, can be a magnificent variable on test scores. Rater variability has been shown to demonstrate itself in many different ways. The main differences, as Brown (2005) asserted, could be variation in their interpretations of the rating scale, level of severity, impressions towards test takers (halo effect), their gender differences as well as test takers' background knowledge. Among these factors, raters' and test takers' gender differences, as proposed by

Hughes (2011), have been presumed to affect the validity of oral performance assessment since they influence both raters' biases in rating and test takers' attitudes towards the rater with whom they take the oral test.

Literature Review

Sources of Rater Variability

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 (Van Moere, 2012). Most importantly, it is now well proved that raters vary regarding their severity in assessing test takers' performance (Winke, Gass, & Myford, 2012). In a study, In'nami and Koizumi (2016) found differences in raters' behaviors depending on various groups of test takers and the type of task in use. Linacre (1989) had already referred to this factor as *bias*.

Besides the above-mentioned factors which influence raters' ratings, during the last 15 years researchers shifted their focus to the features of raters that may influence their ratings (e.g., Brown, 2005; Eckes, 2005; Nakatsuhara, 2011). Among them, according to Lim (2011), oral language assessment and rating experiences are the variables that have attracted the most concentration. In other words, one of the growing concerns in raters' scoring is whether they have been adequately trained or have had enough expertise in assigning accurate scores (Davis, 2016). Assigning accurate scores depends on the experiences that a rater has, cognitive factors, the characteristics of the rating criteria and the rating environment. It is well proved that such differences cause raters to vary with respect to the degree of strictness with which they assess test takers' oral performances (McNamara, 1996).

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 (Winke, Gass & Myford, 2012). According to Cumming (1990) experience refers either to the period of time the rater has been rating or to the amount of rating the rater has done, whereas expertise refers to the raters whose ratings are consistently good. Although experiences and expertise are related issues, they are different in a way that experience may or may not lead to expertise.

A variety of studies on experienced and inexperienced raters' performances have indicated higher inter-rater consistency following training (Ahmadi & Sadeghi, 2016; Attali, 2016; Cumming, 1990). Commonly, in all the studies, extremely severe or lenient inexperienced raters benefited from the training program thus modified their rating behavior. In a study by Ahmadi and Sadeghi (2016) on the effect of rater training on raters' consistency in scoring test takers' oral language proficiency, the consistency of inexperienced raters improved much more after training compared to experienced raters. Some studies have found that inexperienced and experienced raters used a different rating approach to evaluate their students' performances. For example, Attali (2016) found that experienced trained raters used an approach, commonly known as the Funnel Model (a process in which raters score all performances on one feature and then categorize them on the basis of other features), to guide their judgments.

Raters' and Test Takers' Gender Effect

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. According to Aryadoust (2016), differences between male and female raters regarding their interviewing styles can be accounted for as gender effect on assessment. Besides, he argues that raters' behaviors of either gender might change according to whether they are interviewing test takers of the same or opposite gender. After all, in both above-mentioned cases, such gender effects can seriously influence test takers' oral performance assessment and the scores they are awarded either severely or leniently. Furthermore, test takers' performances may change based on their own gender. This in a way means that test takers' performances can be influenced either positively or negatively according to the gender of the rater who is interviewing them. Consequently, the quality and quantity of oral performance and assessment can be affected in two ways, either by differences between male and female raters in scoring or by test takers' genders when performing on oral tasks.

There have been several research studies, which have investigated the possible impact of gender on the scoring of test takers' oral performance by the raters (Buckingham, 1997; Maria-Ducasse & Brown, 2009; O'Sullivan, 2000; Porter, 1991; Porter & Shen, 1991; Winke, Gass, & Myford, 2012). A majority of these studies have demonstrated some sort of gender impact on test scores although the impact is not similar. Some of these studies have shown that test takers were awarded higher scores

with male raters (e.g., Porter, 1991), whereas some others have shown the reverse in that test takers were awarded higher scores with female raters (e.g., O'Sullivan, 2000; Porter & Shen, 1991). Besides, even some studies have demonstrated some interactional effects between the test takers and raters in a way that test takers were awarded higher scores when they were interviewed by the raters of the same gender (Buckingham, 1997). Aryadoust (2016) argues that male and female conversational differences might affect raters' scoring differences. This can depend on whether they are paired with the test takers of the same gender or not. Moreover, this issue can influence raters' severity estimates by enhancing or worsening test takers' test performance, and of course the final outcome of the test. O'Sullivan (2000) studied the impact of raters' gender differences in assessment. In his study twelve Japanese test takers were rated once by a male and once by a female rater. The comparison of scores, except for one case, indicated that the test takers were awarded higher scores when rated by a female rater regardless of the test takers' genders. Moreover, the outcome showed that, the test takers tended to produce more accurate speech when rated by female raters. Finally, the quality of performance reached the highest when both the rater and the test takers were females. However, such finding is in contradiction with that of Caban's (2003) study in which he investigated the oral performance of thirteen male and female Arab test takers as they were rated once by a male and once by a female rater. The outcome indicated that test takers were awarded higher scores when they were rated by male raters. In another study by Winke, Gass, and Myford (2012), investigating the potential source of bias and gender, they found 11% of bias in total interactions in which the raters were biased to the test takers of their own genders. Such contradictory outcomes should better be interpreted on the basis of personality factors of test takers (Van Moere, 2012).

One further consideration of the effect of gender in the assessment of test takers' oral performance by the raters is the test takers' own gender. There have been some research studies investigating the possible effect of

gender on test takers' oral performance rating. A majority of these studies reported some kind of effect on test scores, but with a small effect size (e.g., Aryadoust, 2016; O'Sullivan, 2000). However, in a study by O'Loughlin (2002) investigating the effect of rater gender on rating, the outcome revealed neither any significant effect of gender on their rating nor any behavioral differences between the two. In this study, O'Loughlin (2002) investigated the oral performance of sixteen Asian male and female test takers on an IELTS interview test in Australia on two different occasions, once with a female and once with a male interviewer. The results showed that gender did not have any significant effect on their performance. However, in some exams, like the IELTS, in which the interviewer also acts as the rater, this could cause an uncertainty of whether any assessment fluctuation is due to the test takers' gender, or raters' gender or even a combination of both. He, then, concluded that such contradictions in terms of the effect of gender on oral performance assessment are common, which let us identify the role of context as the source of such contradictions and not necessarily gender in oral assessment. In a rather similar study, Lumley and Sullivan (2005) investigated the effect of test takers' gender on their oral task performance in Hong Kong and found that female test takers outperformed male ones although the difference was not significant. O'Sullivan (2002), in another study on interactional speaking test focusing on Japanese raters' reactions to oral placement tests, found no significant difference between the reactions of male and female raters to students' oral productions.

However, a number of other studies have identified significant difference between male and female test takers' performances in which, according to their researchers, the differences derive from gender effects. Porter (1991), in an interpersonal speaking assessment study, focused on the gender and the personality of sixteen Arab test takers. The outcomes revealed a significant influence of gender on raters' rating biases in a way that test takers (specifically female ones) benefitted more when they were rated by the raters of the same gender. Similarly, Porter and Shen (1991)

studied the effect of gender on raters' scoring behavior and found it significant. The 28 male and female test takers participating in their study scored higher when they were rated by female raters. They further argued that this might be due to raters' cultural or behavioral differences. Fairly consistent with these studies is the one done by Buckingham (1997), who investigated the effect of gender on test takers' performance. It is noteworthy to indicate that unlike Porter and Shen (1991), here the interaction between raters' and test takers' genders were also accounted for. The outcome showed that test takers got higher scores when they were paired by the raters of the same gender. The reason for such contradiction in outcome might be due to various analytical procedures used in the two studies. While Porter and Shen (1991) used t-tests, Buckingham (1997) used an ANOVA, which enables the researcher to investigate inter-variable interactional impacts.

Some studies reported that male raters tend to rate second language oral performances more favorably for male test takers (e.g., Xi & Mollaun, 2011), whereas others have found that females were scored higher when rated by female raters (e.g., Eckes, 2005; O'Sullivan, 2000). In a comparison study of test takers' oral and written performances, Eckes (2005) found bias on the side of the raters benefitting female test takers. Female test takers outperformed male ones and were awarded higher scores on their oral performances. In a study by Hyde and Linn (1988), examining the meta-analysis covering 165 studies, they found an overall mean effect size of 0.11, which indicated that females were slightly superior to males in verbal performance. A more specific analysis of effect size revealed a significant difference between them ($t=0.33$, $p<0.05$) for speech production. Thus, the expectation of this part of the current study was that female test takers would outperform male ones when it comes to oral performance.

The Multi-faceted Rasch Measurement (MFRM) introduced by Linacre (1989), which can be done using the computer software FACETS, takes a different approach to the phenomenon of rater variation by not only

investigating rater factors in performance-based language assessment but also providing feedback to the raters on their rating performance (McNamara & Lumley, 1997). Using the MFRM, McNamara and Lumley (1997) found that rater training could establish higher consistency and less bias. In this approach, rater variation is seen as an inevitable part of the rating process, and rather than being an obstacle to measurement, it is considered actually beneficial because it provides enough variability to allow probabilistic estimation of rater severity, task difficulty, and test taker ability using the same linear scale.

Statement of the Problem and Purpose of the Study

Although claims have been made about the relationship between gender and test performance, little research has investigated the effect of test takers' gender differences on the quality of oral language performance as rated by the raters of different genders. In addition, the impact of raters' gender differences on the consistency and severity measures of test takers' oral language assessment is unknown, i.e., whether test takers' oral performances vary when they interview with male or female raters. More recently, a lot of studies which have found gender differences in oral performance were criticized for over-generalizing their outcomes regardless of accounting for situational and contextual factors and the gender of their interlocutors. Besides, measures of male and female rater biases toward the categories of the rating scale are still vague. Consequently, this research was aimed to investigate the impact of test takers' gender differences on their oral performance quality as assessed by the raters as well as identifying any significant effect of raters' gender differences on test takers' performances when they are rated by the raters of the same or opposite gender.

In this study, FACETS was used, to measure test takers' performances when they were rated by the raters of the same and opposite genders. The analysis was performed for the raters of both groups of experienced and inexperienced raters to observe any possible differences caused by

experience. Furthermore, it was used to determine raters' biases in scoring male and female test takers. Besides, FACETS was also employed to measure the impact of raters' gender differences in scoring the test takers' oral performance on each category of the analytic rating scale to observe any hypothetical variation with regard to their biases within each category. Based on the above-mentioned goals, the following research questions were formed:

RQ1: Is there any significant difference between male and female test takers' oral performance ability?
RQ2: Do raters' gender and experience differences have an impact on the scores they award to test takers of the same or opposite gender?
RQ3: Is there any significant difference between male and female raters' biases toward each category of the rating scale?

Methodology

Participants

A convenient sample (the test takers were selected randomly among only those strata who were studying at intermediate, upper-intermediate and advanced levels) of 300 adult Iranian EFL students, including 150 males and 150 females, ranging in age from 17 to 44 participated in the study as test takers to take the Community English Program (CEP) speaking test. The students were selected from intermediate, upper-intermediate, and advanced levels studying at the Iran Language Institute (ILI) in Tehran.

Twenty Iranian EFL teachers, including 10 males and 10 females, ranging in age from 24 to 58 participated in this study as raters. These raters were undergraduates and graduates of English language related fields of study, teaching in different universities and language institutes. In order to fulfill the requirements of this study, the raters were classified into two groups of experienced raters and inexperienced ones to investigate the similarities and differences among them and the likelihood

of one group rating the candidates differently compared to the other when it comes to the rating of male and female test takers.

In order to search for rater participants for the present study, a background questionnaire, adapted from McNamara and Lumley (1997), eliciting the following information including (1) *demographic information*, (2) *rating experience*, (3) *teaching experience*, (4) *rater training* and (5) *relevant courses passed* was given to the raters. Based on the above-mentioned method of rater classification, raters were divided into two levels of expertise on the basis of their experiences outlined below.

A. Raters who, according to McNamara and Lumley (1997), had less than two years or no experience in rating and receiving rater training, and had less than five years or no experience in teaching and had passed less than four core courses (pedagogical English grammar, phonetics and phonology, second language acquisition and second language assessment) related to ELT (English Language Teaching) major. Hereinafter these raters are named 'NEW'.

B. Experienced raters who had over two years of experience in rating and receiving rater training, and over five years of experience in teaching and had passed all the four core courses plus at least two selective courses (any ELT course in addition to the above-mentioned ones) related to ELT major. Hereinafter these raters are referred to as 'OLD'.

A more important reason for choosing these groups of expertise is to investigate any differences between experienced and inexperienced raters in terms of how they approach the task of oral assessment and how they are affected by the rating process. It is noteworthy to indicate that in order to eliminate rater expectancy effect, the raters and rater groups were not informed of the existence of two various groups and any similarities and differences between the two. Table 1 displays the summary characteristics of the raters participating in the study.

Table 1.
Rater Background Characteristics

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

Instruments

The following instruments were used in this study:

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. CEP is an international and valid oral proficiency test used to evaluate students' speaking ability. The purpose of the speaking test is to measure the extent to which second language speakers can produce meaningful, coherent, and contextually appropriate responses to five different tasks including Description, Narration, Summarizing, Role-play and Exposition tasks.

Task 1 (*Description Task*) is an independent-skill task, which reflects test takers' personal experience or background knowledge to respond in a way that no input is provided for it (Bachman, 2004). On the other hand, tasks 3 (*Summarizing Task*) and 4 (*Role-play Task*) reflect test takers' use of their listening skills to respond orally. In other words, the content for the response is provided for the test takers through listening, short or long. For tasks 2 (*Narration Task*) and 5 (*Exposition Task*) the test takers are required to respond to pictorial prompts including sequences of pictures, graphs, figures and tables. The aim for the use of the above-mentioned five tasks is to enable the researcher get a picture of test takers' oral proficiency in various contexts and genres (Luoma, 2004).

The TOEFL test. It was already mentioned that the test takers were selected from intermediate, upper-intermediate and advanced levels at the ILI; however, considering the sole educational level could not be a valid criterion for classifying learners into different proficiency levels. Thus, in

order to make sure that the test takers taking part in this study were *not* at the same level of language proficiency, a TOEFL test was given to make sure whether there was a significant difference between them or not. The TOEFL test type used in this study is the one obtained from the TOEFL iBT 24 Mock Tests designed by the Educational Testing Service (ETS).

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. The purpose of using an analytic rating scale was to assess test takers' oral performance on the basis of a valid and reliable scoring rubric and to identify how well male and female raters use the rating scale categories, based on the given descriptors, systematically and without bias. Test taker's task performance was assessed using the ETS (2001) analytic rating scale for this study. In ETS (2001) scoring rubric, individual tasks are assessed using appropriate criteria including *fluency*, *grammar*, *vocabulary*, *intelligibility*, *cohesion* and *comprehension*. Each of these criteria is accompanied by a set of seven descriptors.

Procedure

The 300 students studying at intermediate, upper-intermediate, and advanced levels at the ILI in Tehran were randomly selected to take a sample TOEFL (iBT) test including only listening and speaking skills to make sure that they were not at the same level of oral language proficiency and that there was a significant difference among the three groups of intermediate, upper-intermediate and advanced levels. The speaking section of the test was rated by the researcher of the study along with two other colleagues of his who were all authorized by the ETS as examiners. The outcome of the data analysis through running an ANOVA test indicated a significant difference between the three groups of test takers oral ability ($F_{(2, 297)} = 2197.362$, $p<0.01$). This confirms that the test takers

were at various levels of oral proficiency as measured by statistical analytic tests.

Prior to collecting any data from the test takers, the raters' background questionnaire was given to the raters to fill out before starting to run the test tasks and collect data. The aim of having the raters fill out the raters' background questionnaire sheets was to enable the researcher to classify them into the two groups of rating expertise (i.e., NEW raters and OLD ones). Afterwards the tasks of the CEP test were run one by one. It was planned that the room in which the oral test would be held was arranged with comfortable chairs set around a coffee table, a camcorder was placed at a far side of the room, pre-focused on the place in which the test was going to be held so that no camera operator would be needed. This was all done in order to reduce the number of potentially distracting factors in the validity of the study. Each test was recorded on a video tape for future assessment analysis. The students were given 60 seconds to prepare their responses. Each assessment on the tasks lasted approximately around 12 minutes. The responses were recorded in a MP3 format and saved on a CD for the raters to score. The test takers were not allowed to talk to each other about the oral test and the tasks in the exam setting and were supervised by a member of the research coordinators while waiting for their turn to take the test. It is again reiterated that individual tasks are assessed using appropriate scoring criteria including fluency, grammar, intelligibility, vocabulary, cohesion and comprehension consisting each criterion of a set of 7 descriptors ranging from 1 to 7 using the ETS scoring rubric.

All the raters who participated in this study were given one week to submit their scorings, based on the six band analytic rating scale, to the researcher. Moreover, the videotaped recordings of the oral assessment settings were awarded to the raters to assist them observe aspects of oral performance including metalinguistic behaviors and body language.

Data Analysis

Raters' scoring performance was investigated through multiple observations. Quantitative data (i.e., raters' scores based on an analytic scoring rubric) were collected and analyzed with MFRM during the scoring sessions. The scoring patterns of male and female raters for the two groups of experience (NEW & OLD) were investigated each time they scored test takers' oral performances of each gender using the analytic rubric. The quantitative data were compared (1) across male and female raters of the two expertise groups to investigate the raters' ability cross-sectionally at each rating point, and (2) within each rater group for each gender type to investigate any rating difference among the raters of the same experience level.

In the present study, MFRM was performed using the FACETS software program to examine both individual rater and rater group scoring patterns. A 6-facet model was used including the facets of *test taker* (test takers' ability), *test taker gender* (the gender of test takers), *rater* (raters' severity/leniency), *rater expertise* (experienced/inexperienced), *rater gender* (the gender of raters), and *scale criterion* (categories of the analytic rating scale). The facets tables used in the study represent the partial credit model from the facet variable map. Each test taker was rated by all the raters. The FACETS was run to investigate both test takers' proficiency measures when they were rated by raters in terms of their gender and their expertise levels, and raters' biases regarding both their expertise and their genders when they rated test takers' oral proficiency.

Results

The effect of gender on test performance could be investigated on two viewpoints. On the one hand, test takers' genders could be a source of score variation when they are rated by the raters of the same gender. On the other hand, raters' own gender differences may cause score variation when they award higher scores to the test takers of the same genders as their own.

To achieve this goal, the researcher performed a gender bias analysis by estimating the measure for both male and female test takers and male and female raters with separate analytical measures to test the null hypothesis that neither raters nor test takers had any specific bias in the test scoring and test performance respectively. The outcome demonstrates whether the raters scored specific test taker genders more or less severely than the other one, or whether the test takers of a particular gender group outperformed the other when they were rated by raters of a specific gender group.

Having analyzed the data, the FACETS variable map representing all the facets was obtained. In the FACETS variable map, presented in Figure 1, the facets are positioned on a common logit scale which facilitates interpretation and comparison across and within the facets in one report. The figure plots test takers' ability, test takers' gender, raters' severity, raters' gender, rater group expertise, and scale criterion difficulty. The amount of variance of the facets of the study indicates the relative effect of test takers' ability, raters' severity, raters' gender, rater group expertise, and scale categories on test scores.

The *first column (Logit Scale)* in the map displays the logit scale. It acts as a fixed reference frame for all the facets. It is a true interval scale that has the property of equal distances between the intervals (Schaefer, 2008). Here, the scale ranges from 4.0 to -4.0 logits.

The *second column (Test Taker)* displays estimates of test takers' proficiency. Each star represents three test takers. Higher scoring (more competent) test takers appear at the top, whereas lower scoring (less competent) ones appear at the bottom. Here, the range of the test takers proficiency ranges from 3.81 to -3.69 logits; thus making a spread of 7.50 with respect to test takers' ability. It is noteworthy to indicate that no test taker was identified as misfitting – not having an infit mean square value beyond 0.6 and 1.4 logits according to Myford and Wolfe (2004) – thus none of them was excluded from data analysis. Test takers' oral

proficiency mean score was measured 20.43 for the five tasks and the standard deviation was measured 3.07.

The *third column (Test taker gender)* displays the test takers' genders in terms of their competency in oral performance measures. It should be indicated that the genders appearing at the top are more competent than the ones at the bottom. Here, males (logit value = 0.39) were more competent than females (logit value = 0.25), thus making a spread difference of just 0.14 logit value. This difference logit value indicates that there seems to be not much difference between male and female test takers' oral proficiency measures.

The *fourth column (Rater)* displays raters regarding their severity or leniency estimates in rating test takers' oral proficiency. Since there were more than one rater scoring each test taker's performance, raters' severity or leniency scoring patterns can be estimated. This renders raters' severity measures. In this column, each sign represents one rater. Severer raters appear at the top, whereas more lenient ones at the bottom. Ideally, raters should differ very little from each other in the levels of severity they have showing that the criteria for assigning measures is used equally by the raters. Rater OLD8 (Female) (Severity measure: 1.72) was the severest rater and rater NEW6 (Female) (severity measure: -1.97) was found to be the most lenient rater. Besides, in this phase, OLD raters, on average, were rather severer than NEW raters who tended to be more lenient than the OLD ones. Here, raters' severity estimate ranges from 1.72 to -1.97 logits, which makes the distribution of rater severity measures (logit range = 3.69) much narrower than the distribution of the test takers' proficiency measures (logit range = 7.50), in which the highest and lowest proficiency logit measures were 3.81 and -3.69, respectively. This demonstrates that the effect of individual differences on behalf of raters on test takers was relatively small (Schaefer, 2008). Raters, as shown in the figure, seem to have spread equally above and below the 0.00 logits.

The *fifth column (Rater gender)* displays the raters' genders in terms of severity measures. Obviously the genders appearing at the top are

severer than the ones at the bottom. Here, females (logit value = 0.33) were slightly severer than males (logit value = 0.27), thus making a spread difference of just 0.06 logit value. This difference logit value indicates that there seems to be very little difference between male and female raters' severity measures. This column has the least variation compared to the other columns in which all the elements are gathered around the mean.

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

Columns eight to *thirteen (Rating scale categories)* display the six-point rating scale categories used by the raters to score the test takers' oral performances. The horizontal lines across the columns are the categories threshold measures which indicate the points at which the probability of getting the next rating (score) begins. The figure shows that each score level was used although there was less frequency at the extreme points. Most important of all, the FACETS variable map tells us that, as an example of this case, a test taker whose proficiency estimate is 2.0 logits on the logit scale is likely to get a 5.0 in Intelligibility on an *average-difficult* category when s/he is assessed by an *average-severity* rater, or similarly, a test taker whose proficiency estimate is 3.0 logits on the logit scale is likely to get a 6.0 in Fluency on an average difficult category when s/he is assessed by an average-severity rater.

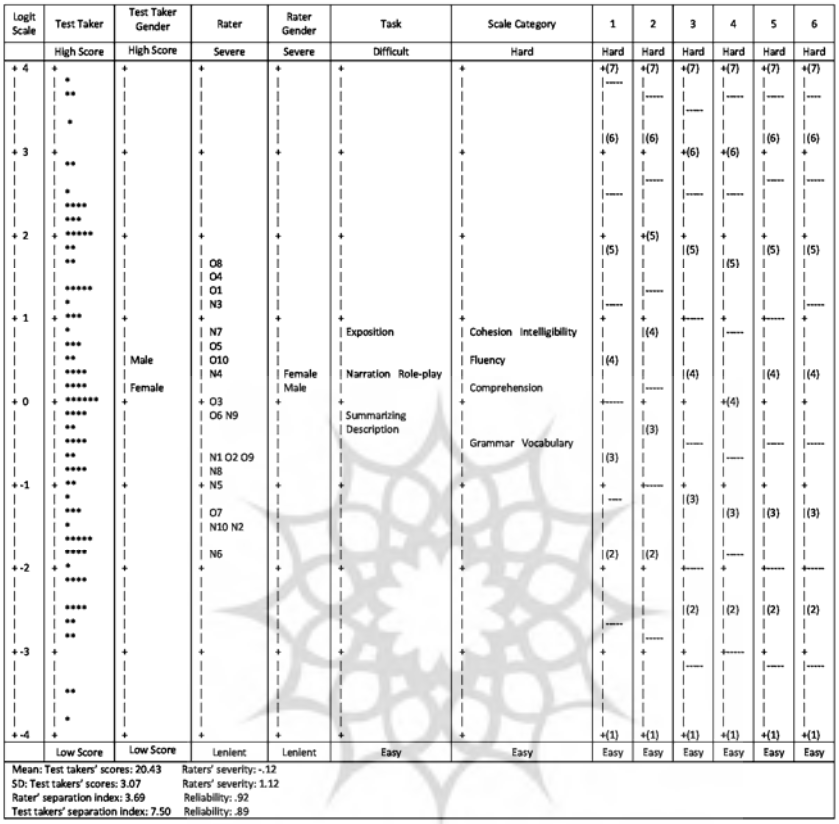

Figure 1. FACETS variable map

First, the effect of test takers' gender differences on their oral performance ability was investigated.

RQ1: Is there any significant difference between male and female test takers' oral performance ability?

The average proficiency measures, displayed in Table 2, demonstrate that the two groups of test takers were close together regarding their speaking proficiency as rated by the raters. The average proficiency measure for male test takers was 0.73 (SE = 0.02) and for the female test takers 0.82 (SE = 0.02). The difference between the average proficiency measures was 0.09 logits. In order to compare the average proficiency measures for the two subgroups of test takers, male and female, an

independent samples *t*-test was run (McNamara, 1996). The average proficiency measure for male test takers was not significantly different form that of the female ones, ($t_{(298)} = 2.87$, $p = 0.089$).

Moreover, according to Myford and Wolfe (2004), the differences of less than 0.30 logits between the performances of test taker groups are not considered as significant; thus, the test takers participating in this research study, of both genders, were statistically at equal proficiency levels. Besides, in order to further ascertain that there is no significant difference between male and female test takers in terms of their proficiency measures, 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 test takers with regard to their oral proficiency measures when they were rated by raters ($X^2_{(1, N=2)} = 5.76$, $p>0.05$).

Table 2.

Test-takers Proficiency Measure Report

Test taker groups	Observed raw score	Observed count	Observed raw score average	Average proficiency (logits)	SE
Male	6976	150	2.4	0.73	0.02
Female	7085	150	3.1	0.82	0.02
Mean	7030		2.75	0.77	
SD	7.07		0.49	0.06	
Fixed (all same) chi-square: 5.76, *df*=1, *p*=0.24					

Second, the effect of raters' gender and expertise differences on the scores awarded to male and female test takers was examined through the following research question:

RQ2: Do raters' gender and expertise differences have an impact on the scores they award to test takers of the same or opposite gender?

Table 3 displays the FACETS analysis, as the partial credit model of the map of variables, of the two groups of test takers, male and female, the quality of their performances and thus, the scores they got based on whether they were rated by the raters of the same gender or not. It also presents the analysis of test takers' performances as they were rated by NEW and OLD raters to indicate to what extent raters' expertise affects

raters' gender differences in their oral proficiency assessment. The *Obs-Exp average score (column six)* is what is known as bias size but calculated on the basis of raw scores (Myford & Wolfe, 2004). This is obtained by subtracting the observed count from the expected raw scores divided by the observed counts.

Table 3.

FACETS Analysis of Male and Female Test Takers' Performance

Test Taker	Male	Male	Female	Female	Mean	SD	Fixed (all same) Chi-square: 8.45, *df*=1, *p*=0.17 Gender separation index: 1.14 Reliability index:0.37
Rater	OLD	NEW	OLD	NEW			
Obs. Raw Score	3504	3472	3567	3518	3515.2	39.50	
Exp. Raw Score	3368.7	3306.6	3442.8	3419.7	3384.4	60.43	
Obs. Count[a]	150	150	150	150	300	0.00	
Obs –Exp score (Logit)	1.35	1.65	1.24	0.98	1.30	0.27	
Bias Size[b]	-0.09	-0.12	-0.08	-0.07	0.02	-0.09	
Z-score	-0.47	-0.58	-0.43	-0.34	-0.45	0.09	
SE	0.02	0.03	0.02	0.02	0.02	0.00	
t	4.28	3.86	4.41	4.56			
df	149	149	149	149			
r	0.74	0.62	0.76	0.71			
Rasch Logit Measure[c]	0.16	0.15	0.15	0.17			
p	0.14	0.21	0.20	0.16			

[a] The count of the number of rated oral performances that contributed to the observed raw score
[b] The higher the bias size, the more biased the rater will be towards the test taker groups
[c] The average level of severity that the raters displayed when rating test takers' oral performances of each group

The *bias size (column seven)* is the amount of bias measures in logit values for male and female test takers. Positive measures show severity and negative measures show leniency of test takers performances and the scores they were awarded when they were rated by male and female raters of the two expertise groups. The mean bias value (in logits) was measured -0.09, thus the test taker' groups display more than half a logit value above or below the mean logit value (between -0.59 and 0.41), which would be considered as either too severe or too lenient (Eckes, 2015; McNamara, 1996). In this respect, both male and female test takers were identified within the acceptable range of severity and leniency measures. This indicates that male and female test takers were awarded similar scores on their performances regardless of the fact that they were rated by male or female raters.

Z-scores (column eight) indicate the amount of test takers' biases of each gender to oral performance assessment when they are rated by the raters of various expertise levels. ***Bias*** is the difference between expected and observed ratings of the obtained data, which is then divided by its standard error to achieve the Z-score (Wright & Linacre, 1994). According to McNamara (1996), Z values between ±2 are considered as the acceptable range of biasedness, thus any values above or below the given Z score are considered either too positively biased, or too negatively biased. Here the data showed that both male and female test takers were within the acceptable range of biasedness with respect to their bias measures regarding raters by whom they were rated. In other words, male and female test takers' oral performance scores were not affected by raters' gender differences.

SE (column nine) displays the standard error of bias estimation. The small amount of SE provides evidence for the high precision of measurement. The *t* and *p* values indicate whether the obtained bias

differences between the test taker gender groups are significant or not. The obtained bias size is the size of test takers' bias to the raters' gender groups and the *t*-value represents each bias size. The outcome demonstrated no significant difference between male and female test takers when they are rated by male and female raters of either expertise group.

The *r* value *(column twelve)* indicates point biserial correlation coefficient found between male and female test takers rated by male and female raters of each group of expertise. The outcome showed relatively similar correlation between the test takers performances. The outcome according to Cohen's table of effect size was identified as much higher than the typical value showing high extent of significant correlation. Moreover, as shown in Table 2, the *Chi-square* results indicated that there was no significant difference between the two groups of test takers with regard to their oral performance ability when rated by the raters of the same gender ($X^2_{(1, N=2)} = 8.45$, $p>0.05$).

The *separation* and *reliability index* of male and female test takers performances were measured 1.14 and 0.37, respectively, reflecting that test takers could not be separated based on their oral performances regarding gender differences. The separation index (1.14) displays that test takers were rather in a single level of proficiency. The reliability index (0.37) also represents that the separation of test takers into various levels was rather vague.

The results of the bias interaction analysis did not show any bias of either group of test takers to neither male nor female raters of either group of expertise. In all cases, the interaction effect size (Bias) was very small (less than 10%) indicating that the test takers' biases did not incorporate substantive differences in their performance ability regarding the raters' gender differences. In other words, the mean bias logit measured 0.03; therefore, the test takers with bias logit values beyond half a logit value of the mean logit would be regarded as being too severe/lenient (Wright & Linacre, 1994). In the above data analysis, no test takers displayed a significant bias value beyond the acceptable range; thus, no gender group test takers were treated too severe or too lenient by the raters.

Additionally, a similar FACETS analysis was performed in order to observe any probable amount of bias to any category of the rating scale, which might have been hidden when the overall scores were observed and analyzed. This issue was questioned in the following research question:

RQ3: Is there any significant difference between male and female rater biases toward each category of the rating scale?

Table 4 displays the bias interaction between raters' gender type and each category of the rating scale. It investigates whether male and female raters were biased in their ratings of any of the categories of the rating scale or not. The outcome of the table vividly indicates that there was no significant difference between male and female raters in their scorings of test takers using each category of the rating scale. In this respect 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 regarding their bias to each category of the rating scale ($X^2_{(1, N=2)} = 2.19$, $p>0.05$). In other words, the raters did not show any significant bias to any of the categories of the rating scale (See Table 4).

Table 4.

Raters Gender-Scale Category Bias Interaction

Rater groups	Average raw score	Bias in rating scale categories: Cohesion Bias size (Z)	Sig.	Intelligibility Bias size (Z)	Sig.	Fluency Bias size (Z)	Sig.	Comprehension Bias size (Z)	Sig.	Vocabulary Bias size (Z)	Sig.	Grammar Bias size (Z)	Sig.	SE
Male	1.9	1.34	NS	1.12	NS	0.52	NS	0.17	NS	0.11	NS	-0.08	NS	0.03
Female	2.4	0.74	NS	0.89	NS	0.33	NS	-0.07	NS	-0.23	NS	-0.16	NS	0.05
Mean	2.15	1.04		1.00		0.42		0.05		-0.06		-0.12		0.04
SD	0.35	0.42		0.16		0.13		0.16		0.24		0.05		0.00

Fixed (all same) Chi-square: 2.19, $df = 1$, $p>0.05$

Discussion

Regarding the first research question which dealt with the difference between male and female test takers' oral performance abilities, the outcome of data analysis showed no significant difference between the two genders. This finding is in line with that of O'Loughlin (2002) who found neither any significant effect of test takers' genders on their rating nor any behavioral differences between the two. Such finding is also parallel with the one found by Lumley and Sullivan (2005) and O'Sullivan (2002) who in the investigation of the effect of test takers' genders on their oral task performance did not find any significant difference between male and female test takers in their oral performance abilities. Nevertheless, this finding is in contrast with those of Aryadoust's (2016) and O'Sullivan's (2000) who, in separate studies, found a significant gender effect on test scores, though with a small effect size. More contradictory findings were observed through the comparison of the finding of this study indicating the insignificant effect of test takers' gender impact on their oral performances with the ones found by Nakatsuhara (2011), Porter (1991), Porter and Shen (1991) and Buckingham (1997) who found a significant gender influence on raters' rating biases in a way that test takers (specifically female ones) benefitted more when they were rated by the raters of the same gender. The insignificant effect of test takers' gender differences in their oral performance ability contradicts with those of Xi and Mollaun's (2011) and Eckes's (2005) who found that test takers demonstrate better performances when they are paired by the raters of the same gender.

One reason for the contradiction between the outcome of this study and those of the above-mentioned studies might be the difference in the type of statistical analyses that were employed. The mentioned studies mostly used *t*-test and ANOVA, which restrict the researcher to observe the interactional effect of few variables on one another, whereas this study used the MFRM which enabled the researcher to meticulously observe a combination of the interactional effects of all the variables used in the study. Another reason that could possibly justify the differences between

the findings of this study and other studies as explained above, could be the possible cultural differences that might have existed between the present context and the context of the cited research studies. In other words, O'Sullivan's (2000) study was on Japanese raters and test takers, Porter (1991) focused on Arab participants, and the study by Xi and Mollaun (2011) involved Indian raters. Consequently, although the findings of this study were parallel with those of some previous research, the contradictory outcomes with a number of other studies might be due to cultural differences and contextual variations. The interesting point, however, is that although the context of the current research lacked a coeducational setting, the findings revealed no gender bias in raters' scorings. Likewise, the test takers were not disturbed or influenced by the gender of the raters with whom they were paired. Thus, it might be the case that test takers' oral ability overrules the effect of educational setting; nevertheless, further research is needed to provide evidence supporting this assumption. Yet, another justification might be the educational background of the raters who participated in this study, which might have been different from the raters of the other studies.

Regarding the second research question focusing on the effect of raters' gender differences and experience differences on the scores they awarded to the test takers of the same or opposite gender, the outcome of the study did not show any biasedness of either group of raters to neither male nor female test takers. This indicates that there is no interaction between the effect of raters' gender and expertise on their bias toward assessing male and female test takers' oral performances. Although there is rather little previous work related to this finding, O'Loughlin (2002), in an investigation of raters' bias toward test takers' oral assessment on the IELTS test, found bias on the side of inexperienced raters in assessing either male or female test takers' oral performances. Previous research (e.g., Lim, 2011; Winke, Gass, & Myford, 2012) has demonstrated differences in terms of bias between experienced and inexperienced raters showing that by virtue, experienced and inexperienced raters, seemed to

have different perceptions from one another and of course in their judgment of test takers' performances which originated from their idiosyncratic characteristics. Examples of such differences can be tolerance of mistakes (Van Moere, 2012), consistency measures (O'Sulivan, 2002), and severity or leniency (Maria-Ducasse & Brown, 2009). Nevertheless, this does not seem to affect their judgments regarding test takers' gender differences in oral performance assessment.

Regarding the third research question dealing with differences between male and female rater biases to the rating scale categories, the outcome of data analysis demonstrated no significant difference between male and female raters in their scorings of test takers based on the rating scale categories. Some previous studies have demonstrated that raters show halo effect when they face problems using rating scales. For example, Sawaki (2007) stated that when raters cannot identify certain aspects of rating scales, they resort to more global and holistic use of rating scales. However, this issue mostly arises from their variability in interpreting the meaning of each criteria and its relevant descriptor. On average, inexperienced raters tended to be more lenient in a majority of the rating scale categories than experienced ones. Accordingly, the findings of this study are in line with the study by Fall, Adair-Hauck and Glisan (2007) who found that inexperienced raters were significantly more lenient in their ratings of coherence and fluency, and the one by Davis (2016), who found that experienced raters were significantly harsher in their ratings than inexperienced raters in assessing speaking ability. Meanwhile, it must be noted that the obtained outcomes are fairly contradictory to that of Brown (2005) who found that inexperienced raters tended to be severer than experienced ones with respect to the scoring of test takers' pronunciations who, mostly, overfitted the model. However, this study once again indicated that such differences in terms of the use of the rating scale categories have nothing to do with raters' gender differences. Such differences in raters' interpretations of scale categories can be resolved through appropriate rater training programs, and the remaining differences

regarding bias measures could probably be attributed to the other intervening variables.

Conclusion and Implications

The findings of this study suggested that gender, either on account of the raters or the test takers, does not have any significant impact, on the one hand, on the performance ability of the test takers and on the other hand, the biasedness of raters' scoring patterns. In other words, test takers' gender differences do not affect the quality and quantity of their oral performance. Besides, raters' scoring is not affected by whether they are paired with the test takers of the same or opposite gender since the data did not show any bias caused by gender differences. The outcomes of this research provide further evidence on the sameness of male and female rater performances in rating. Therefore, there will be no excuse on the side of decision makers to exclude the raters of either gender from rating test takers' oral performances. The findings of this study are both in line and in contradiction with previous research studies which found both a nonsignificant and a significant gender effect on test takers' performances and raters' biases in rating – although the significant effects had come up with inconsistent directions. Such contradictions between the outcome of this study and the previous ones might be due to the insignificant nature of gender impact in the CEP oral test. Such insignificant effect of gender among test takers and raters reduces the salient effect of gender in subsequent research.

Another reason for the contradiction in the findings can be attributed to social perspectives, cultural backgrounds, and contextual issues, which may cause differences in performance and assessment, rather than to gender. Such social and contextual issues can be the country where the test is administered and the social identities and background of test takers and raters. It is clear that gender differences are inevitable in assessment in the field of applied linguistics. This suggests that gender might be mistaken for social and contextual factors, and as a result in various assessment

contexts, different outcomes are achieved. This possibility implies the necessity of researching the effect of social and contextual factors on rating bias. In summary, one cannot simply predict when gender has a significant effect on oral assessment, and this issue could be plausibly true for both raters and test takers. However, it is noteworthy to indicate that one cannot always predict when gender plays a significant role in oral assessment and when it does not. With regard to this study, care must be considered when generalizing the outcome due to the relatively small number of raters participating in the study.

Gender also was shown to have an insignificant effect on raters' bias in the use of rating scale categories which shows that differences in the use of scale categories is a matter of their various interpretation of rating scale descriptors rather than their gender differences. Likewise, differences between experienced and inexperienced raters were shown to be gender neutral. That is, common differences between inexperienced and experienced raters, and the typical leniency of inexperienced raters in scoring was shown to be independent of gender differences.

The findings of this study have a number of practical implications in education. Firstly, raters' gender is a factor, which does not affect test takers' oral performances. Consequently, unlike previous findings which reiterated the deletion of gender effect to reduce stress and increase the validity of assessment, this study suggests that there is no need to neutralize the impact of gender for a more valid and reliable assessment. Secondly, since inexperienced raters are more economical than experienced ones, they cost less for decision makers in large-scale assessment. Therefore, instead of charging a bulky budget on experienced raters, decision makers can allocate the budget to running more efficient training programs. Nevertheless, this research did not address the impact of gender on other language skills (e.g., writing) and other modes of oral language assessment (e.g., semi-direct). So, further research is required to investigate the effect of test takers' gender differences and raters' gender bias accounting for these skills and modes of assessment.

References

Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. *Language Assessment Quarterly, 13*(4), 341-358.

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

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99-115.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Baleghizadeh, S., & Gordani, Y. (2012). Core units of spoken grammar in global ELT textbooks. *Issues in Language Teaching, 1*(1), 33-58.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study on their veridicality and reactivity. *Language Testing, 28*(1), 51-75.

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

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

Buckingham, A. (1997). *Oral language testing: do the age, status and gender of the interlocutor make a difference?* Unpublished MA dissertation, University of Reading.

Caban, H. L. (2003). Rater group bias in speaking assessment of four L1 Japanese ESL students. *Second Language Studies, 21*(1), 1-44.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51.

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

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement.* Frankfurt: Peter Lang Edition.

ETS (2001). *ETS Oral Proficiency Testing Manual*. Princeton, NJ.: Educational Testing Service.

Fall, T., Adair-Hauck, B., & Glisan, E. (2007). Assessing students' oral proficiency: A case for online testing. *Foreign Language Annals, 40*(3), 377-406.

Fulcher, G., Davidson, F., & Kamp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29.

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

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104*(1), 53-69.

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing, 33*(3), 341-366.

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.

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

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing, 22*(4), 415-437.

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

Maria-Ducasse, A., & Brown, A. (2009). Assessing paired oral Raters' orientation to interaction. *Language Testing, 26*(3), 423-443.

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

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

Myford, C. M. & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement. *Journal of Applied Measurement, 5*(2), 189-227.

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

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

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

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair task performance. *Language Testing*, *19*(3), 277-295.

Porter, D. (1991). Affective factors in the assessment of oral interaction: Gender and status. In S. Anivan (Ed.), *Current developments in language testing* (pp. 99-102). Singapore: SEAMEO RELC.

Porter, D., & Shen, S. (1991). Sex, status and style in the interview. *The Dolphin*, *21*(2), 117-128.

Sawaki, Y. (2007). Construct validation of analytic rating scales in a speaking assessment: Reporting a score profile and a composite. *Language Testing, 24*(3), 355-390.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

Sunderland, J. (1995). Gender and language testing. *Language Testing Update, 17*(1), 24-35.

Van Moere, A. (2012). A psycholinguistics approach to oral language assessment. *Language Testing, 29*(3), 325-344.

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231-252.

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

Xi, X, & Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Language Learning, 61*(4), 1222-1255.