Teaching English Language, Vol. 14, No. 1, Winter & Spring 2020, pp. 71-90

http://dorl.net/dor/20.1001.1.25385488.2020.14.1.3.0

The Impact of Rater Bias on the Language Performance Assessment Scores of Iranian Foreign Language Teacher Candidates

Ali Hosseini

English Department, Faculty of Paramedical Sciences, Shiraz University of Medical Sciences

Nasrin Shokrpour[1]

English Department, Faculty of Paramedical Sciences, Shiraz University of Medical Sciences

Abstract

Utilizing the scores obtained from a teacher entrance test used in an English language institute as a means of selection, the researchers selected 100 out of 121 female teacher candidates to participate in this study. Furthermore, a reading, writing, and listening test was administered to the candidates to exclude those candidates with low and high proficiency. Based on the results obtained from the tests, the number of participants decreased to 30 and they were requested to come in for oral interviews; they were those who were interviewed twice by two different groups of male and female raters. The results analyzed through correlational analysis and descriptive statistics indicated that the interview scores of the teacher candidates, as measured by the first group of female raters (G_2) was in high correlation with those in class performance (*p*). However, there was no correlation both between the interview scores of the female teacher candidates (G_1), assessed by the male raters and their in class performance (*P*) and the (G_1) - (G_2) pair. It can be concluded that rater bias might have had an effect on the Iranian female teacher candidate's test scores. The subjects were also divided into attractive and unattractive groups and further assessed by the fourth and fifth group of female and male raters to indicate whether female sex appeal affects the test scores or not. The results showed that the mean differences between the AF-AM (attractive female-attractive male) pairs were significant although the mean differences between the NAF-NAM (nonattractive female-non-attractive male) (*P*=➢ *.05, t=1.131)* were not significant.

***Keywords*:** Correlational analysis, Oral Interview, Rater Bias, T-Test

Received on January 27, 2019
Accepted on February 3, 2020

[1] Correspond author: shokrpourn@gmail.com

1. Introduction

Typically learners' test performance is reported in the form of a set of scores. Differences often exist between subgroups of an entire population and are attributable to differences in the ability or construct being measured. However, they may also be wholly or partially caused by the measuring procedure used. When assessing the language of their students, the teachers expect to find different language skills and abilities and assign them different scores. Regardless of the testing method, the reliability of ratings is one of the major controversial issues in language assessment (McNamara, 1996). Put otherwise, scores can be affected by factors outside the ability being measured and this introduces the concept of bias into assessment. Test bias which was of great interest to researchers during the 1970s and early 1980s (Berk, 1982; Jensen, 1980; Reynolds & Brown, 1984) has been examined under two broad lines, namely the psychometric and socio-cultural frameworks In the psychometric approach, the test instrument and student's responses to it are investigated, while in the socio-cultural framework researchers' focus on the performance of the students as part of the overall context in which a student lives and learns (Schellenberg, 2004).

Generally speaking, a test is said to be biased when there are systematic differences in the meaning of test scores based on group membership or when people from different groups who have the same observed score do not have the same standing on the trait of interest (Backman, 1990). Test bias may also be present in cases where a test that is intended to predict some construct results in systematic over- or under-prediction based on group membership. Broadly speaking, there are two models for test bias, namely the mean difference and equal regressions. In the former, the most intuitive definition of bias is observation of a mean difference between the groups, while the latter discusses bias into the context of the interpretation of test scores. In

another rough classification, test bias is defined as the reflection of bias in test meaning and/or use. One type of bias is construct bias which emerges when a test is shown to have different meanings for two groups in terms of the construct it intends to measure. Construct bias is concerned with the relationship of the observed scores to true scores on a psychological test. In cases in which the relationship is systematically different for different groups' test, it is said that the test is biased. Sometimes, construct bias leads to situations in which two groups have the same average true score on a psychological construct but different average test scores. Another type of bias is predictive bias which transpires when test use has different implications for two groups. Based on the previously mentioned assertions, it may be concluded that, like reliability and validity, test bias is a theoretical concept. Therefore, estimating or detecting the degree of test bias is complex if not impossible. In other words, there is no single method to detect test bias any more than there is a one way to estimate reliability or validity. However, it is feasible to determine the degree to which test bias exists.

In a sense, each and every constructed piece of research or study is more or less of some prominence to and influential on the available body of knowledge; nevertheless, some are far more conspicuous. This study sheds some light on test bias while localizing it to Iran. The implications are then assumed to be twofold; on the one hand, scholars, teachers, and experts in the field procure conversancy regarding test bias in countries where gender has a determining role in social affairs, not exceptionally learning and teaching. On the other hand, this awareness gears them with the prerogative to seek a solution to test bias. In accordance with the aforementioned assertions and with respect to the implications of such studies, it is worth mentioning that although the very nature of research is discovering myths and providing answers to hypotheses, in many cases research still fails to reach the

predetermined goals not because of lacking a comprehensive data base, but for the reason of its impropriety to the context. The former point may also be taken as further evidence for the significance of the present study. Furthermore, the majority of the conducted studies in Iran do not precisely focus on rater bias, rather they consider test bias. This latter point may be taken to contribute to the innovation of this study.

The aim of this study was to challenge the so long held belief that females are better language learners than males and also investigate the possibility that rater bias may have influenced Iranian female students' test scores. In Iran, female students usually obtain higher scores on language tests. This has led some (teachers, parents, etc.) to believe that females are better language learners compared to males. However, in practice this is not the case, especially with male raters. Male students are seen to perform at the same level as female students, despite their lower marks of language tests, especially on oral interviews and situations where the female student is known to the rater. These and many more led the researchers to construct this study as means of providing answers to the following questions:

1. How does test bias on the part of the male raters affect the obtained scores of a test of oral proficiency for female English foreign language students?
2. In what ways do female Iranian foreign language students' biological characteristics cast an appealing impression on male raters so that they unintentionally assign extra points to them?

2. Literature Review

2.1 Test Bias

According to social sciences, during evaluation inherent judgments reflect the thing or person being assessed and the built-in biases. In other words, judgments are affected by perceptual goggles. The prominence of the fact becomes even more evident when it is understood that the effects of the

rater's perceptions introduce highly subjective factors that make many evaluations more or less inaccurate Language teaching and learning as a social phenomenon is no exception. Misinterpretation, miscommunication and/or failure to perform or misunderstanding the illocutionary acts of a language may have devastating consequences. Thus, all judgments and evaluations need to be as valid and reliable as possible, which is more easily said than done.

Many language test bias studies have found unexpected interactions between rater judgments and other facets not related to the test takers' performance. For example, Wang (2010) investigated rater agreement in China and concluded that raters showed differences in interpreting and applying rating criteria. Native and nonnative raters were presumed to show even greater differences in rating (Wang, 2010). In a similar study, Zhang and Elder (2011) found no significant differences between the scores assigned by natives and nonnative raters to oral performance of a group of examinees. However, the two groups were found to differ in the way they weighed various aspects of oral proficiency construct.

In line with other studies, Son (2010), outlining recommendations for reducing bias in elicited imitation (EI) administration and creating a registration tool for collecting the raters' background information, discussed the rater bias, stressing the bias attributable to raters and test takers' language backgrounds. In a similar way, Caban (2003) investigated language background and educational training to determine whether these factors have an effect on the raters' assessments of Japanese medical students in a controlled oral interview. The interpreted statistics show variation though not seemingly as a result of the factors being investigated.

Likewise, Zhao (2017) investigated the raters' performance and focused on not only how the raters' second language proficiency level interacted with the

examinees' first language, but also if the raters' teaching experience had any effect on their scores. The findings indicated that extensive rater training can be quite effective: there was no significant effect of either the raters' familiarity with examinees' L1, or raters' teaching experience on the scores. However, even after training, the raters still exhibited different degrees of leniency/severity.

2.2 Bias in Oral Interviews

Recent research into oral language interviews has indicated that interviewers vary considerably from each other regarding their test behavior. Such variability includes the amount of support they give to candidates, the amount of rapport they establish with them, and the extent to which they follow the instructions relevant to their role (e.g., Brown & Hill, 1996; Lazaraton, 1996; McNamara & Lumley, 1997; Morton, Wigglesworth & Williams, 1997; Young & Milanovic, 1992). An increasing number of studies have focused on rater variability in performance-based assessment of L2 ability. Raters have been proved to differ with regard to the severity of their evaluation of examinees' oral proficiency and can produce a broad range of scores. Raters have also been found assigning the same score to disparate performances or disparate scores to the same performance (Orr, 2005). Studies indicate that even if high degree of agreement exists between the raters, this does not by any means guarantee similar judgment. Stated otherwise, the same score may mean different things to different raters (Ang-Aw & Meng Goh, 2011; Johnson & Lim, 2009).

Another strand of research addresses what Sunderland (1995) refers to as the gender effect. Here, differences in male and female interviewers' styles can be viewed as a potential gender effect. The behavior of interviewers of either gender may vary according to whether they are paired with a male or female candidate. In both cases, the possibility exists that the gendered

behavior of the interviewer will influence the outcome of the test by either strengthening or undermining the candidate's performance. There have also been a number of recent studies which have examined the possibility of a gender effect in the rating of candidates by their interviewers in oral interviews. Most of this research reveals some kind of gender effect on test scores, though interestingly, the effect is not always the same. Some studies report that test-takers scored more highly with male interviewers (e.g., Locke, 1984; Porter 1991a, Porter 1991b), while others report higher scores with female interviewers (e.g., O'Sullivan, 2000; Porter & Shen, 1991). An interaction effect between the genders of the interviewer and interviewee has also been reported (Buckingham, 1997). That is to say, the candidates achieved a higher score when paired with an interviewer of the same gender. By virtue of their very inconsistency, these findings appear to support more recent thinking about the shifting and unstable nature of gender in spoken interaction.

2.3 Test Bias in the Iranian Context

Contrary to the above studies, the research conducted in Iran does not specifically focus on rater bias; rather, it considers test bias or focuses on rater bias and tests of writing. For instance, Farhady (1979) claimed that students with different linguistic and educational backgrounds might perform differently on discrete-point and integrative tests. The findings indicated that male and female students with different educational backgrounds perform better or worse on one or the other type of test. Regardless of the degree of reliability, validity and practicality, some undesirable biases exist which may be due to variables which are not directly relevant to the underlying constructs of the tests. He concluded that the concepts of discrete-point and integrative tests need serious attention in the development of second language proficiency tests.

Karami (2011), applying Rasch's model, investigated the presence of DIF among male and female examinees taking the Tehran University English Proficiency Test (UTEPT). The results revealed that 19 items functioned differentially for the two groups. Only 3 items, however, displayed DIF with practical significance. This implies that the existence of DIF may be interpreted as impact rather than bias (p. 27).

In another study, Alaei, Ahmadi, and Sabourian Zadeh (2014) investigated whether or how personality traits, analytical/holistic scores, and genre interact in EFL writing assessment. The findings showed no significant relationship between the raters' traits and the holistic scores assigned to each genre. However, significant correlations were identified between analytic scores given to each individual component of scoring rubric and the raters' traits. The authors conclude that making raters aware of their personality traits can help them to find out the sources of their biases, and their tendencies to respond in certain ways to texts. In a similar vein, Khoshsima & Roostami Abusaeidi (2015) investigating if English Major and non-English major teachers differed in their perception of the construct of oral proficiency while assessing the learners' L2 oral proficiency. The findings indicate that while the inter-rater reliability was relatively high for both groups, the English major teaches were generally more reliable during the assessment.

Storehouses have been written on test bias (e.g., Congdon, 2006; Engelhard, 2002; Lumley & McNamara, 1995; Lunz, Stahl, & Wright, 1991; Lynch & McNamara, 1998; McNamara & Adams, 1991; Moon & Hughes, 2005; Nijveldt et al., 2009; O'Neill & Lunz, 1996). However, most of the studies have focused on bias resulting from sex, race, ethnic group membership, social status, or other factors that cause discrimination among different groups in the society, and few have examined bias resulting from an

interaction between raters and some facet concerned with the examinee or test (rater-examinee, rater-rating scale, or rater-task). Therefore, this study was conducted to investigate the possible relationship between rater bias and the Iranian female EFL learner's oral proficiency scores on an oral interview and determine whether the assigned grades are consistent with their level of proficiency or have been affected by rater bias. Although research on this topic may be abundant, what distinguishes the latter from the former is the context. As previously stated, the study localizes the context so that the findings are as close as possible to and meet the needs of the stakeholders in Iran.

3. Method

3.1 Participants

Because of the Iranian-Islamic identity and culture, raters usually refrain from acting and behaving informally towards the people on the opposite sex to avoid the associated negative consequences. Therefore, finding volunteers to freely participate in such studies is very difficult and requires a confidential procedure. Apart from this, the method of participant selection is also seriously in need of being kept confidential to eliminate or decrease the Hawthorne effect on the part of the raters and test takers. As an alternative, the researchers decided to select the participants and collect the data from an English language institute, one of whom was a manager who was an intimate friend of the researcher. The name of the institute and its location is kept confidential so that further conflicts are kept to the minimum. Overall, a number of 136 individuals were enrolled in the study. There were 15 male and female raters (Male = 5, Female = 10) along with 121 female individuals (who were candidates for becoming future teachers in the institute). However, the number of female teacher candidates decreased to 30 as a result of the administration of a reading, writing and listening test. In addition to the

first group of five male raters, five female raters were also selected as secondary raters in order to evaluate the female teacher candidate's performance on an oral interview a second time. After the initial and second assessment stages, the same female teacher candidates actual in class performance were rated again by different groups of five female raters (these group of raters attended class as observers). A fourth and fifth group of male and female raters were also involved in the study. This group of male and female raters was responsible for assessing the individuals in the second stage/phase of the study.

3.2 Instruments

The instruments implemented in this study were tests of reading, writing and listening along with a structured oral interview in which the participants were involved in a conversation with the raters. In order to assign points to the participants' speaking abilities and mark deficiencies in their speech, a check list was used (Appendix). The general structure of the interview was based on Lynch's (1996) interview guide. The questions began with a warm-up phase, eliciting information regarding the use of phatic talk including the student's age, and educational background. In the next phase, more technical questions were asked. A question in this stage might go as: 'Do you consider yourself fit for teaching?' Or 'Why do you like teaching English?' In the third phase, the students were placed in more sophisticated situations and asked to perform in contrary situations. For instance, students during the interview were deliberately told that they were to teach in another city rather than the one they applied for. At the end of the three phases, the individuals were evaluated in terms of accuracy, fluency, and the overall mastery over the target language. The rater then assigned scores to the overall performance of the participants.

3.3 Procedure

The data collection phase was divided into three main stages. The candidates were asked to come in for an oral interview. The first group of male raters assigned scores to the participants' performance regarding accuracy, fluency, and the overall command over the target language, in this case English. The obtained results were then put aside so that test effects were demolished. In the second phase, the female teacher candidates were again interviewed on the previously mentioned aspects, this time by the group of female raters. Then, the results of the first assessment were compared to those of the second assessment and both were compared to the teacher's actual future in class performance. A comparison between the three evaluations revealed any possible correlation between test scores and the teacher's mastery over the oral aspects of the English language. Other things equal, it was assumed that the stronger the correlation between the obtained scores on the two oral interviews and the actual performance of the female teachers, the more reliable the interview scores and the less the effect of rater bias.

In order to investigate the effect of rater bias on the test scores of the female teacher candidates even more, the subjects were further divided into two groups, namely those female teacher candidates that were more attractive (very pleasing in appearance or sound and or causing interest or pleasure sexually) and those who were not. Note that, what we mean by attractive women here is those women or girls who dress not solemnly, are attractive, and employ their feminine characteristics in a seductive way. As with the previously mentioned part of this study, the two groups were asked to come in for an oral interview with a different group of male and female raters. The assessment done by the group of male raters was compared to that of female raters.

3.4 Data Analysis

For the purpose of the study, two sets of analysis were conducted. First, to assess the mean differences between the three sets of scores, descriptive statistics were used. Second, to examine the correlation between the three set of scores, correlational analysis was applied. Moreover, because other factors may also have an effect on the students' English language proficiency, there was an attempt to minimize them to the extent possible by asking for the students' demographic data, examining them and excluding exceptional cases (e.g. those who may have started learning English in an English speaking country) from data analysis. Moreover, for the second part of the present study, the teacher candidates' performance (the two groups) on the first and second assessments was compared using descriptive statistics, a paired sample t-test and an independent sample t-test to identify any existing mean differences between the two sets of scores.

4. Results and Discussion

Tables 1 and 2 present the means (M), standard deviation (SD) and results of the correlational analysis for the teacher candidates' three sets of scores (G_1, G_2, P). The mean of the teacher candidates' score on the oral interview as measured by the first group of male raters (G_1) was the highest among all variables ($\overline{x}$ =85.26), while that of the teacher candidates' score on the oral interview assessed by the first group of female raters (G_2) was lower than those on the oral interview, as measured by the first group of male raters ($\overline{x}$=77.26). As for the teacher candidates actual in-class performance (*P*), the mean score was seen to be closer to the obtained mean score of the female teacher candidates' score on the oral interview which was evaluated by the first group of female raters ($\overline{x}$ = 80).

Table 1
Descriptive Statistics

	$\overline{X}$	SD	N
G_1	85.26	8.75	30
G_2	80.00	7.98	30
Performance	77.26	7.75	30

In addition to the use of descriptive statistics, correlational analysis was also run to evaluate the correlation between the two sets of oral interview scores and the teacher candidates actual in class performance. To this end, the female teacher candidate's performance on the two oral interviews and their actual in-class performance were compared with each other ($G_1$1 -G_2; G_1-p;G_2-p). The results indicated a strong correlation between the G_2-*p* pair (p$<.05$), whereas, there was no correlation between the other two sets of scores (Fig. 3 & 4, & 5).

As indicated by the correlational analysis and descriptive statistics, the performance of the female teacher candidates assessed by the first group of male raters was better. In view of the above, the female teacher candidates may be said to have been assigned higher scores compared to their performance on the same oral interview rated by the first group of female raters. Data analysis indicated mean differences between the assessments of the female candidates on three occasions. Assuming that the reliability and validity indexes are satisfactory, it could be concluded that because the mean score of the second assessment of the same group of female teacher candidates which was measured by the first group of female raters ($\overline{X}$=.80) was lower than the first assessment of the same group of female teacher candidates, measured by the group of male rater ($\overline{X}$ =85.26) and meanwhile it was closer to the obtained mean of the same group of female teacher candidates in class performance ($\overline{X}$=77.26). Rater bias could be said to have had an effect on their scores. In other words, male raters had assigned higher

scores to female teacher candidates based on biological factors rather than the subjects' overall mastery over the oral aspects of English.

Table 2
The Result of the Correlation Test

		G_2	*P*
G_2	Pearson Correlation	1	.951
	Sig. (2-tailed)		.000
	N	30	30
p	Pearson Correlation	.951	1
	Sig. (2-tailed)	.000	
	N	30	30

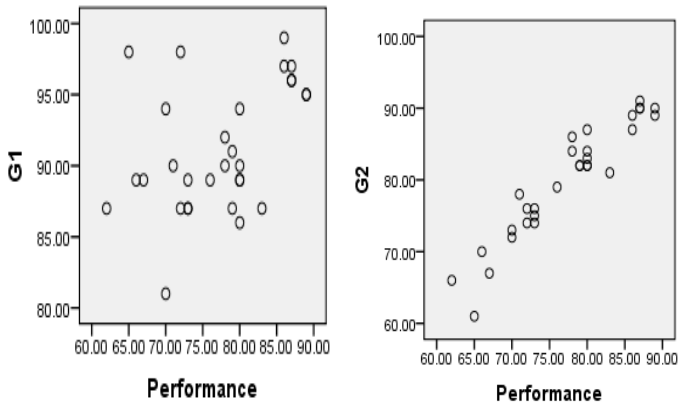

Figure 1. Distribution of the scores

In line with the first part of the study and in order to see whether female Iranian foreign language students' biological characteristics, i.e. beauty, flirting, etc. cast an appealing impression on male raters so that they unintentionally assign extra points to females, the subjects were divided into two groups of attractive (A) and nonattractive women (NA). Both groups were assessed by two groups of male (the AM and NAM groups) and female raters (AF and NAF groups). After the assessment, the obtained results were

analyzed by descriptive statistics, a matched and independent t-test to indicate significant or non-significant mean differences between the AM-AF pair, the NAM-NAF pair (NAM and NAF are indicators of the nonattractive group of women as assessed by the group of male raters and non-attractive women as assessed by the group of female raters respectively) and the AM-NAF pair. As illustrated in Table.3, the mean differences between the AF-AM pair *(P=.000< .05, t=7.1* and the AM-NAF pair *(P=.000< .05, t=1.131)* was significant, but those between the NAF-NAM (*P=> .05, t=1.131)* was not significant. As formerly assumed, the attractive group of female teacher candidates had performed differently on the two occasions, whereas the nonattractive group performed nearly the same on the two interviews; the evidence from data analysis indicated that rater bias had affected the attractive and non-attractive female teacher candidates' performance on the two interviews. Figure 2 shows the distribution of scores for the two groups.

Table 3
Results of the Matched and Independent T-tests

	$\bar{X}$	SD	t-value	Sig
AF-AM	1.97	10.7	7.14	.000*
NAF-NAM	1.53	5.24	1.131	.277
AM-NAF	17.9	3.74	4.790	.000*

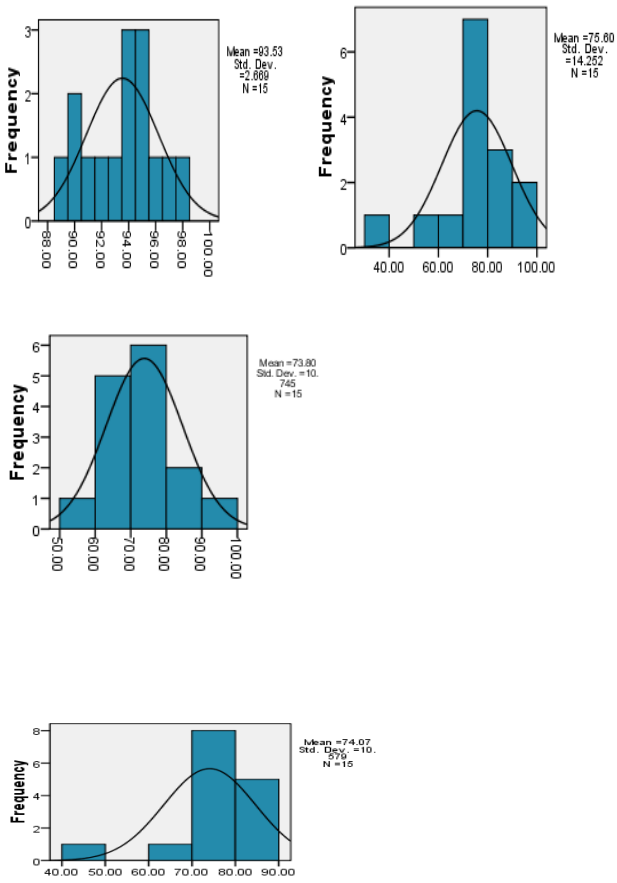

Figure 2. Distribution of scores in the two groups

5. Conclusions

The foot prints of women can be seen on men's decisions and actions throughout history. Many of the fought battles of the past were the mere result of the spell women have casted on men, a fact which has very well been continued to the modern world. Today's educational settings flourish with females who are striving to perform at the full potential for a better future. Considering the competitive job market and the presence of sexism in

the occupational structure, women are left with no choice than to use all the resources at their hand, not least of which being their feminine characteristics. They often dress, talk, and use their body language in such a way to influence other people, especially male bosses, interviewers, managers, and so on. With respect to this and the fact that testing and test performance have always been of importance to teachers, decision makers, and to the whole educational system and that various inner and outer criteria may affect the individuals' overall performance and test score, one of which being test bias, the oral mastery of the same group of teacher candidates to use English in teaching situations was compared on five occasions and rated by five different groups of raters. Significant and nonsignificant mean differences were identified between the assessments. It could be concluded that the first and second groups of male raters were biased towards females so that they assigned points higher than the scores they deserved.

References

Alaei, M. M., Ahmadi, M., & Zadeh, N. S. (2014). The impact of rater's personality traits on holistic and analytic scores: Does genre make any difference too? *Procedia Social and Behavioral Sciences, 98,* 1240–1248.

Ang-Aw, H., & Meng Goh, C. (2011). Understanding discrepancies in rater judgment on national-level oral examination tasks. *RECL Journal, 42*(1), 31-51.

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

Berk, R. A. (1982). *Handbook of methods for detecting test bias*. Baltimore: Johns Hopkins University Press.

Brown, A. & Hill, K. (1996). Interviewer style and candidate performance in the IELTS oral interview. IELTS Australia Reports Round 1.

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 the speaking assessment of four L1 Japanese ESL students. *Second Language Studies, 21*, 1-44.

Congdon, P. (2006). Bayesian Statistical Modeling. West Sussex, UK: John Wiley & Sons, Ltd.

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

Engelhard, G. Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large scale assessments for all students: Validity, technical adequacy, and implementation* (pp. 261-288). Mahwah, N. J.: Lawrence Erlbaum Associates.

Farhady, H. (1979). Test bias in language placement examinations. University of California, Los Angeles.

Jensen, A.R. (1980). *Bias in mental testing*. New York: Free Press.

Johnson, J., & Lim, G. (2009). The influence of rater language background on writing performance assessment. *Language Testing,* 26, 485-505.

Karami, H. (2011). Detecting gender bias in a language proficiency test. *International Journal of Language Studies*, 5(2), 167-178.

Khoshsima, H., & Roostami Abusaeidi, A.A. (2015). English and non-English major teachers' assessment of oral proficiency: a case of Iranian maritime English learners. *Iranian Journal of English for Academic Purposes, 1*(4), 26-36.

Lynch, B. K. (1996). Language program evaluation: Theory and practice. Cambridge: Cambridge University Press.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing,* 13, 15-72.

Locke, C. (1984). The influence of the interviewer on student performance in tests of foreign language oral/aural skills. Unpublished MA project, University of Reading.

Lynch, B. K & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of ESL speaking skills of immigrants. Sage publications.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(1), 54-71.

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1991). The invariance of judge severity calibrations. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

McNamara, T. F. (1996). *Measuring second language performance*. New York: 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*, 142–51.

McNamara, T. F., & Adams, R. J. (1991). Exploring rater behavior with Rasch techniques. Paper presented at the 13th annual Language Testing Research Colloquium, Princeton, NJ.

Morton, J., Wigglesworth, G. and Williams, D. (1997). Approaches to the evaluation of interviewer behaviour in oral tests. In Brindley, G. and Wigglesworth, G., editors, *Access: Issues in Language Test Design and Delivery*. Sydney: National Centre for English Language Teaching and Research, 175-96.

Moon, T. R., & Hughes, K. R. (2005). Training and scoring issues involved in large-scale writing assessments. *Educational Measurement: Issues and Practice*, *21*(2), 15-19.

orr, m. (2005). the fce speaking Test: Using rater reports to help interpret test scores. *System*, 143-154.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System*, 28, 373–86.

O'Neill, T. R., & Lunz, M. E. (1996). Examining the invariance of rater and project calibrations using a Multi-facet Rasch Model. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Porter, D. (1991a). Affective factors in language testing. In Alderson, C.J. and North, B. (Eds.), *Language Testing in the 1990s* (pp. 32-40). London: Modern English Publications.

Porter, D. (1991b). Affective factors in the assessment of oral interaction: gender and status. In Arnivan, S. (Ed.), *Current developments in language testing. Anthology series 25* (pp. 92-102). Singapore: SEAMEO Regional Language Centre.

Porter, D. & Shen Shu-Hung (1991). Sex, status and style in the interview. *The Dolphin*, 21, 117–28.

Reynolds, C. R. & R.T. Brown (1984). *Perspectives on bias in mental testing*. New York: Plenum Press.

Schellenberg, S. J. (2004, February*). Test bias or cultural bias: Have we really learned anything.* Annual meeting of the national council for measurement in education. San Diego, California.

Shawcross, P. (2007, March 12). What do we mean by the washback effect of testing? Retrieved September 2017 from: http://icao.int/icao/en/anb/meetings/ials2/Docs/15.Shawcross.pdf

Son, B. (2010). Examining rater bias: *An evaluation of possible factors influencing elicited imitation ratings*. (MA Thesis).

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

Wang, B., (2010). On rater agreement and rater training. *English Language Teaching,* 3, 108-112.

Young, R., & Milanovic, M. (1992). Discourse validation in oral proficiency interviews. *Studies in Second Language Acquisition*, 14, 403-24.

Zhang, Y., & Elder (2011). Judgement of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50.

Zhao, K. (2017, March). *Investigating the effects of rater's second language learning background and familiarity with test-taker's first language on speaking test scores.* Retrieved from https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=7256&context=etd

Appendix

Fluency	Accuracy	Intonation
She was seen to be quite fluent but in the absence of native- like pronunciation, while ample hesitations and pauses could be identified.	In this regard, the female teacher candidate was seen to perform pretty well as long as she was not required to talk about something in detail. As an instance, if the subject was asked about recipes, she would comfortably talk about the ingredients but she faced problems in talking about the method of putting them together.	The assessed female teacher candidate's speech lacked this feature more than others. A clear example of the former point was seen in question formation. The rater had difficulty making out whether the produced sentence was a question or not because the rising and falling patterns could not be identified.
Overall performance		
The overall speaking ability of the female teacher candidate for survival purposes was satisfactory; however, the candidate was not suited for engaging in a long conversation. Note that her speech was full of improper hesitations and pauses.		
Fluency Score	Accuracy Score	Intonation Score
25	42	23