Copyright

by

Yi-Wen Su

2014

The Report Committee for Yi-Wen Su

Certifies that this is the approved version of the following report:

The Impact of Rater Characteristics on Oral Assessments of Second Language Proficiency

APPROVED BY

SUPERVISING COMMITTEE:

Supervisor: ____

Diana Pulido

____

Elaine K. Horwitz

The Impact of Rater Characteristics on Oral Assessments of

Second Language Proficiency

by

Yi-Wen Su, B.A.

Report

Presented to the Faculty of the Graduate School

of the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

May 2014

The Impact of Rater Characteristics on Oral Assessments of Second Language Proficiency

by

Yi-Wen Su, M.A.

The University of Texas at Austin, 2014

SUPERVISOR: Diana Pulido

This literature review sets out to revisit the studies exploring impact of rater characteristics on language oral assessments. Three categories of raters' backgrounds: occupation, accent familiarity, and native language are identified and will be addressed respectively in the following sections. The results showed that no consensus regarding raters' occupational background, linguistic background and native-speaker status on examiners' rating has been found so far. However, this review will highlight the current testing situations, bring up limitations from previous studies, provide implications for both teachers and raters, and hopefully shed light on future research.

Table of Contents

1.0 Introduction

1.1 What is Assessment?

All educational programs must at some point examine learners’ knowledge and abilities; that is, they must assess them. Assessment is a process that involves collecting information and interpreting results so as to evaluate a certain construct of a learner’s ability or skill. Assessments are conducted to improve educational programs’ effectiveness and to track students’ learning and developmental outcomes. A large number of studies have delved into examining validity in testing and assessment, exploring whether a test “measures accurately what it intended to measure” (Hughes, 1989, p. 22). Other studies have looked into the reliability of the test or, in other terms, “desired consistency of test scores.” Crocker and Algina (1986, p. 105) state “whenever a test is administered, the test user would like some assurance that the results could be replicated if the same individuals were tested again under similar circumstances.” The validity and the reliability of a test, therefore, are crucial to give an effective and authentic interpretation of one’s ability.

1.2 What is Oral Assessment?

Oral proficiency in a foreign language is at the very core of communicating effectively. Pronunciation, considered one of the most essential criteria of oral proficiency, includes individual sounds, pitch, volume, speed, pausing, stress, and intonation. Unlike traditional tests that collect fixed responses and are assessed by computers (e.g., multiple-choice tests), oral assessment needs a human rater to listen to speech samples and then award scores.

1.3 The Importance of Raters

In the complex world of language assessment, second language (L2) oral performance assessment always involves raters' subjective ratings. The presence of raters might exert influence on test takers' scores (Lumley & McNamara, 1995) and on the process of rating. Speaking, therefore, becomes the most difficult language skill to assess reliably (Luoma, 2004). The term *rater variability* adds a new dimension to the process of assessment, making the monitoring of reliability and validity even more crucial. *Rater variability* or *rater effect* is defined as "construct-irrelevant variations associated with rater characteristics" (Huang, 2013, p. 771) rather than with the examinees' actual performance or ability. It is critical to the reliability and the validity of speaking assessments (Bachman et al., 1995; Kunnan, 2000). It also has crucial and consequential impacts on decision-making processes, particularly in high-stakes testing situations (Schaefer, 2008). Some testing institutions have tried to reduce the impact of rater variability. The most common procedure that examination boards use to ensure the reliability of their scoring is rater training (Luoma, 2004). Rater training allows raters with different characteristics to practice applying the same standards to assess. In addition, rater training helps raise inter-rater reliability, which is the degree to which raters agree with each other when rating the same performances. That is, the higher the inter-rater reliability is, the more consistent the raters' ratings are.

However, there is very little research on the effectiveness of current rater training programs as well as rater effects obscuring the construct; what there is has yielded conflicting results. Therefore, research examining the validity and fairness of oral performance is still needed.

2.0 Characteristics of Second Language Raters

Rater effects exhibit themselves in various forms. *Rater severity* is one of the most prevalent rater effects in performance assessment. It occurs when raters consistently assign either too harsh or too lenient grades when compared with other raters (Lumley & McNamara, 1995; McNamara, 1996). Apart from rater severity, raters may apply different rating criteria, attend to different linguistic or nonlinguistic features of the performance and weight judging criteria differently. Therefore, the same performance might result in varying ratings, or the same ratings may stem from different reasons (Brown, Iwashita, & McNamara, 2005). In addition, rater background variables, such as raters' rating experience (Cumming, 1990; Weigle, 1994, 1999), occupation (Barnwell, 1989; Brown, 1995; Hsieh, 2011; Huang, 2013; Liu, 2011; Lumley1998; Powers & Stansfield, 1985), trained or untrained (Ang-Aw & Goh, 2011; Barnwell, 1989; Hsieh, 2011; Huang, 2013), accent familiarity (Carey, Mannell & Dunn, 2011; Winke, Gass & Myford, 2012; Winke & Gass, 2013) and first language(s) (Baitman & Campos, 2012; Brown, 1995; Caban, 2003; Kim, 2009; Xi & Mollaun, 2009; Zhang & Elder,2011) may also influence how raters determine their ratings. In order to assign grades impartially, raters should not "impose their own values to the extent that they are aware of them" (International Language Testing Association, 2005, p. 2), which might compromise the interpretation of the scores. Hence, this literature review sets out to revisit the studies exploring impact of rater characteristics on oral assessment. Three categories:

raters' occupation, accent familiarity, and native language are identified and will be addressed respectively in the following sections. This review will also highlight the current testing situations, bring up limitations from previous studies, provide implications for both teachers and raters, and hopefully shed light on future research directions.

2.1 Occupation

The first focus of rater characteristic is raters' occupations. Occupational experts are assumed to have a good command in their professional fields, whereas well-trained raters are assumed to have mastered the language. Whether the occupational expert can rate language performance and the trained rater can grade occupational knowledge are in question. Although It is common practice to use trained language raters for the TOEFL iBT speaking test, the ACTFL oral proficiency interview, the IELTs oral interview and other high-stakes oral proficiency tests, a question emerges as whether or not trained language raters are able to make reliable judgments concerning candidates' abilities to use language in more specialized contexts. As oral occupation-specific tests increase and demand more raters (Liu, 2011), a crucial question is whether language-trained raters or field practitioners should be determining who is qualified for specific occupations. Many researchers have tried to compare the perceptions of language-trained raters and occupational practitioners in oral assessments of occupation examinations (Barnwell, 1989; Brown, 1995; Hsieh, 2011; Huang, 2013; Liu, 2011; Lumley1998; Powers & Stansfield, 1985). However, previous studies yielded conflicting results concerning the effect of raters' occupational knowledge in oral occupation-specific tests. Below is a review of these studies.

Powers and Stansfield (1985) delved into whether occupational experts and their clients made similar judgments as language trained raters in the context of the Test of Spoken English, a test of general English proficiency. They compared 27 nurses' and 26 patients' ratings in audio tapes of 36 test candidates wishing to practice in nursing with trained TSE raters. Results showed moderate levels of agreement between occupational practitioners, their clients and trained language raters.

Barnwell (1989) examined whether linguistically naïve native speakers graded differently from language trained raters in the ACTFL scale. Fourteen naïve native Spanish speakers were recruited to rate four oral interviews tapes. The data collected from the raters were compared with the ratings made by two ACTFL-trained interviewers. The results revealed that the native speakers were consistently stricter in their evaluations than the ACTFL-trained raters.

Brown (1995) explored the effect of raters' professions and native language backgrounds in the context of the Japanese Language Test for Tour Guides. Thirty-three raters including native and non-native Japanese speakers were recruited and divided into two groups based on their occupation backgrounds: Japanese teachers and tour guides. Both groups of raters rated 51 test candidates' interview videotapes on two broad aspects:

linguistic skill and task fulfillment. Linguistic skill was assessed based on a scale from 1-6. Task fulfillment was also assessed on a scale of 1-6; assessors assigned grades based on the quality of the candidate's interaction in terms of its appropriateness for guide-client interaction. The results showed that tour guide raters were slightly harsher than teacher raters for linguistic skill. As for task fulfillment, tour guide raters were slightly more lenient than teacher raters. However, the raters did place difference importance on graded items. While teachers were more conservative on grammar and expression, vocabulary and fluency, tour guide raters marked pronunciation more severely. Overall, Brown's analysis showed that the two groups of raters despite their professions demonstrated consistency in their ratings.

Lumley (1998) examined the impact of raters' occupations on rating in the context of the Occupational English Test (OET), a test for the medical profession. Ten ESL raters trained in the assessment of the OET and nine doctors with extensive overseas experience were selected for the study and rated 20 audio recordings of role plays. Lumley compared the extent of agreement in holistic ratings between the two groups of raters. The finding revealed that although significant variations existed amongst raters of each group, there was a reasonable agreement between the two rater groups.

Liu (2011) investigated the effects of different levels of raters' occupational knowledge on the assessment of English specific purpose (ESP) oral performance. Liu divided 360 raters who held different levels of finance or accounting qualifications into three occupational knowledge groups: a good mastery group, basic understanding group, and no understanding group. Raters rated 30 Chinese students’ ESP oral performance on eight types of speaking tasks using both holistic and analytical scales. Data were analyzed by ANOVA and interviews were conducted to substantiate the quantitative results. The finding revealed that raters’ level of occupational knowledge had varying impact on the ESP oral assessment, depending on the type of task and the rating scale that was used in the study.

Hsieh (2011) set out to explore the differences between ESL teachers and linguistically naïve American undergraduates in rating International Teaching Assistants’(ITAs) oral proficiency. Three aspects of rater severity and rater orientations were investigated: overall proficiency, accentedness, and comprehensibility. Thirteen experienced ESL teachers and 32 linguistically naïve undergraduates rated 28 ITAs speech samples derived from the Speaking Proficiency English Assessment Kit (SPEAK), a semi-direct test used as the ITA screening tool at the university where the research was undertaken. Three holistic rating scales were used. The raters could optionally provide written comments that justified their grading. Results of the study first suggest that rater backgrounds had a minimal impact on rater severity in oral proficiency ratings. However, it shaped the raters’ perceptions of accentedness and comprehensibility. ESL teachers were found to be more lenient in rating accentedness and comprehensibility. Secondly, the study identified a great number of factors that raters paid attention to while evaluating the examinees’ speech, such as linguistic resources, phonology, fluency, content, global assessment and nonlinguistic factors. Lastly, the undergraduate raters appeared to evaluate the examinees' oral proficiency more globally while the ESL teachers tended to rate more analytically by attending to different linguistic features of the speech samples.

Huang (2013) set out to investigate the impact of raters' accent familiarity and their profession on their ratings of oral language proficiency. Three groups of raters who varied in their accent familiarity with Chinese and experience of teaching ESL/EFL were recruited: unfamiliar non-teachers, familiar non-teachers and familiar teachers. Each group consisted of 22 raters and all raters assessed 26 speech samples derived from the TOEFL iBT speaking test. Each rater assigned a holistic score on Overall Proficiency and three analytical ratings on Foreign Accents, Grammar and Vocabulary, and Content, and completed a survey reporting their perceived impact of their own background on ratings and evaluative features that they used for rating. Although the results showed no statistical significance in the two effects among the three groups, the familiar teacher rater group did report they felt their teaching experience assisted their rating.

To summarize, despite the fact that some studies found similar results indicating no significant difference between occupational practitioners and trained language raters (Huang, 2013; Lumley, 1998; Powers & Stansfield, 1985), other studies produced conflicting findings showing one group of raters was stricter than the other (Barnwell, 1989; Brown, 1995; Hsieh, 2011). Additionally, Liu (2011) found that varying levels of occupational knowledge also had differential impact among vocational experts. Hence, with the divergent findings, it is hard to draw a conclusion on the impact of raters' occupation on their ratings. In addition to the different results, these studies were not without limitations. First, the small sample sizes (Barnwell, 1989; Brown, 1995; Hsieh, 2011; Huang, 2013; Lumley, 1998; Powers & Stansfield, 1985) made the results not generalizable to a larger group of raters. A larger-scale study therefore, is needed. Secondly, only a few professions with one or two studies in each field were examined, such as education (Barnwell, 1989; Hsieh, 2011; Huang, 2013), medicine (Lumley, 1998; Powers & Stansfield, 1985) and tourism (Brown, 1995) fields. Future research should include more varied professions so that the results could be applied in other occupational contexts. Thirdly, there might be other characteristics that the researcher overlooked in the study. For example, the undergraduate raters' former personal experience with ITAs might have clouded their ratings (Hsieh, 2011).

2.2 Accent Familiarity

Another focus of rater characteristic is raters' accent familiarity with test takers' first language. Many studies found that raters may better comprehend the speech produced by test takers whose native languages are more familiar on some level (Munro, Derwing, & Morton, 2006). Studies have indicated that listeners adapt to foreign-accented speech when repeatedly presented with it, leading to higher processing speed (Clark & Garrett, 2004), and increased accuracy in sentence recognition (Bradlow & Bent, 2008). Individuals may be familiar with a foreign accent because they are speakers or learners of that language or because they interact with people who have that particular foreign accent. Length and intensity of the exposure may vary, which may lead to different types and levels of accent familiarity. Furthermore, studies have shown that listeners may favor or downgrade certain foreign accents (Major, Fitzmaurice, Bunta, & Balasubramanian, 2002). Questions emerge of how linguistic experience shapes raters' perception and whether raters' accent familiarity leads to rater bias. Previous studies yielded discrepant results of raters' linguistic background being a potential source of rater bias (Carey, Mannell & Dunn, 2011; Huang, 2013; Winke & Gass, 2013; Winke, Gass & Myford, 2012; Xi & Mollaun, 2009). The review below outlines the studies that have investigated this issue.

Xi and Mollaun (2009) set out to investigate whether Indian raters, after being trained and certified, were able to accurately and consistently rate examinees with mixed first language backgrounds, particularly Indian examinees, in the context of the speaking section of the Test of English as a Foreign Language Test (TOEFL iBT). The effectiveness of a special training process designed for raters was examined as well. Twenty-six trained and certified raters who were bilingual speakers of Indian languages and English were randomly divided into two groups and received training in Mumbai, India. While the first group received regular rater training that was similar to that of U.S. operational TOEFL raters, the other group received the regular training plus special training designed to instruct them in how to score native Hindi speakers' English speech samples. Rater feedback surveys were given to collect qualitative data. Using correlation and inter-rater reliability statistics, findings showed that with training, the Indian raters performed as well as the U.S. operational TOEFL raters in scoring both Indian and non-Indian examinees. In addition, qualitatively, the raters reported that the special training assisted them in rating Indian examinees more consistently and boosted their confidence in scoring Indian examinees.

Carey, Mannell and Dunn (2011) examined the impact of the amount of exposure a rater has to non-native English accents on oral assessments. Ninety-nine IELS OPI raters from five geographically dispersed centers located in different countries (India, Hong Kong, Australia, New Zealand and Korea) were recruited. The raters were mostly native English speakers in the four countries except in India. Three authentic IELTS oral interviews conducted in Hong Kong, Korea and India were audio-recorded, and each test candidate was from a different L1 background, which were Chinese, Korean and Indian English respectively. The raters assigned grades based on an OPI rating form. A questionnaire was used also to elicit the raters' demographic information and their level of familiarity with the interlanguages of the three candidates. The findings demonstrated a significant association between the raters' familiarity of test takers' interlanguages and the scores the raters assigned. That is, the results indicated that the rater was more likely to award a higher score when the rater was familiar with the speaker's English accent. In addition, the results revealed that the location of the test centers also impacted the scores assigned to the three candidates. Significantly higher scores were awarded to the candidate at their L1 country than at the other four centers.

Winke, Gass and Myford (2012) investigated whether accent familiarity leads to rater bias. Winke et al. (2012), operationalizing familiarity differently from Xi and Mollaun (2009) and Carey et al. (2011), defined it as having learned the test takers' L1 as a foreign language in the past. They recruited 107 examiners who varied in the level of accent familiarity, education background and the way they acquired (heritage learners) or learned the L2 and had them rate 432 TOEFL iBT speech samples from 72 test takers. The raters were L2 speakers of Spanish, Chinese, or Korean, while the test takers comprised three native-speaker groups of Spanish, Chinese, and Korean. The researcher analyzed the data using a multifaceted Rasch measurement approach. Results revealed that L2 Spanish raters were significantly more lenient with L1 Spanish test takers, as were L2 Chinese raters with L1 Chinese test takers.

Winke and Gass (2013) conducted a separate qualitative study based on the quantitative data collected from a previous study (Winke, Gass & Myford, 2012). As mentioned earlier, the researchers investigated the effect of raters' knowledge of test takers' first language (L1) on the process of rating oral speech. Twenty-six trained raters who were learning a second language of Mandarin, Korean, or Spanish rated English speech samples from 72 English learners who came from Mandarin, Korean and Spanish L1 background. The raters' processes of rating were video-recorded, and then the raters watched the videos of themselves rating and discussed their rating processes. The authors utilized software package QSR NVivo 8 to conduct a qualitative analysis. Eight themes emerged: (1) test-taker's accent, (2) test-taker's L1, (3) rater's heritage status, (4) affect (how the rater felt), (5) test-taker's voice, (6) task, (7) scoring difficulty and (8) technical problems. Results indicated that some raters attended to test-takers' accent and L1. In addition, raters who were heritage language learners discussed how their personal heritage status influenced their comprehension of the speech samples from the same L1 speakers. The author concluded that raters' sensitivity to test-taker accents led themselves to rate the speech in a biased way, compromising test reliability.

Huang (2013) (discussed in the occupation section above) addressed the issue of raters' familiarity with speakers' accents and their profession of teaching English as a second/ foreign language (ESL/EFL). Although the results showed no statistical significance in raters' accent familiarity among the three groups, many raters reported that their accent familiarity with Chinese affected their rating decisions.

In conclusion, while most studies (Carey, Mannell & Dunn, 2011; Winke, Gass & Myford, 2012; Winke & Gass, 2013) showed raters' bias from accent familiarity with test takers’ L1, Huang (2013) and Xi & Mollaun (2009) found the opposite. Hence, the effects of accent familiarity on raters’ score assignments are not clear. In addition, when it comes to trained raters, the two studies that employed trained raters and investigated the effects of accent familiarity on scoring (Carey et al., 2011; Xi & Mollaun, 2009) came to differing conclusions. In Xi and Mollaun (2009), although accent familiarity did not directly affect raters’ scoring, raters did show concern when rating examinees with the same L1; and the authors reported that accent familiarity should be dealt with during rater training. In Carey et al. (2011), accent familiarity was found to significantly affect ratings. As a result, no conclusive results were drawn.

Furthermore, the studies all had some weaknesses. First, the raters employed in the studies differed a lot in their backgrounds. For example, raters selected for the study (Winke et al., 2012) varied substantially in age (18-61) and educational and occupational backgrounds and were imbalanced in gender distribution (30 male; 77 female); these variables might have, to some extent, confounded the results. Considering the gender factor for an example, some studies reported test-takers scored more highly with female raters (e.g.,O’ Sullivan, 2000). Future research should recruit raters from similar backgrounds and employ raters that better match the qualifications of operational raters in large-scale language-testing-rating programs. Another limitation is that the studies had different operational concepts of “accent familiarity”. Some studies included raters who had learned test takers' L1 as a foreign language (Huang, 2013); or lived in the test takers' country (Carey et al., 2011); other studies consisted of raters who were native speakers of the test takers' L1 (Xi & Mollaun, 2009) while some studies were comprised of raters who were bilingual and thus were familiar with test takers' L1 (Winke et al., 2012; Winke & Gass, 2013). Future research should define clearly what type of raters they are looking at to make their results more reliable by comparing the same type of raters. Furthermore, raters' learning examinees' L1, raters' living in examinees' country, raters' sharing the same L1 as the examinees' and raters' degree of bilingualism in examinees' L1 could all be looked at separately as different rater backgrounds. The third limitation is particularly within Huang's (2013) study. As the author admitted, the study would have been strengthened by having a complete 2x2 design. That is, the rater group was incomplete without non-familiar teachers. The result of the familiar teacher group, accordingly, was hard to attribute to raters' familiarity with the language or to their teaching experience.

An additional limitation is that most of the studies looked primarily at quantitative data (Carey et al., 2011; Huang, 2013; Winke et al., 2012; Xi & Mollaun, 2009). Despite collecting qualitative data from raters' written comments, the studies still did not provide information of how well each rater was actually rating. Future research should value qualitative data by conducting interviews with the raters and include information on how raters assigned grades to particular speech samples. Lastly, since multifaceted Rasch analysis could be helpful in identifying biased raters; in order to uncover the nature or awareness level of rater biases, these biased raters should be targeted and interviews should be conducted to help gain insight into rater biases. To conclude, since research has found that accent familiarity might be a potential factor that inadvertently affect a rater's decision-making process; further research is needed to provide a deeper understanding of this rater characteristic.

2.3 Native-Speakerness

The last focus, an especially heated debate for language testing researchers, is the role that native speakers (NS) and non-native speakers (NNS) play in language assessment. Traditionally, a native English speaker (NES) is defined as a speaker of Standard English (Davies, 1999), and the normative system of native speakers has long been assumed in language teaching and testing. Unsurprisingly, those large-scale, high-stakes tests such as the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS) based their assessments using native English-speaking ability (Lowenberg, 2002). However, several challenges emerged against the NS norm. The first challenge is the definition of native speaker. It is difficult to draw an absolute and clear boundary between native and non-native speaker; therefore, the NS norm has been recognized as a myth (Davies, 2003). Secondly, cited from Zhang & Elder (2011, p. 33) is "the World Englishes (WEs) challenge to the NS model (Lowenberg, 2002)". That is, all English speaking countries have the right to develop their distinct localized forms of English. The World Englishes (WEs) movement has begun to document and describe different varieties of English, for example Hong Kong English (Joseph, 2004) and Singapore English (Lowenberg, 2002) (Zhang & Elder, 2011, p. 33). Moreover, new varieties of English, such as 'Japlish' (Stanlaw, 2004), and 'Chinglish' (Lowenberg, 2002) are said to be developing in countries where English is widely used but is not an official language. These Englishes can potentially function as standards in their own right. Besides, a strong WEs view maintains that to impose international English on users of WEs may be discriminatory against non-native English speakers whose legitimate local uses may be treated as errors (Lowenberg, 2002). Lastly, as English becomes a language of international communication and the number of non-native speakers of English in the world outnumbers that of native speakers (Yano, 2002), whether NS norm should continue to be the only acceptable standard is in question (Taylor, 2006). It is therefore crucial to examine whether non-native English speaker (NNES) raters are as consistent and as competent as native English speaker (NES) raters. With little research and contradictory findings (Baitman & Campos, 2012; Brown 1995; Caban, 2003; Kim, 2009; Xi & Mollaun, 2009; Zhang & Elder, 2011), more research is needed to provide a clearer picture of the rating behavior of NS and NNS. The review below examines these issues.

Brown (1995) (discussed in the occupation section above) explored the effect of raters' native language backgrounds in the context of the Japanese Language Test for Tour Guides. She found minor differences in terms of "harshness" between the two groups of raters. First, native speakers were found to be stricter than non-native speakers. Second, native speakers showed more diverseness in their harshness than did non-native speakers. Third, non-native speakers were significantly stricter with items such as politeness and pronunciation than were native speakers. As for the task fulfillment assessments, non-native speakers were slightly more lenient than were native speakers. In addition, non-native speakers were generally less consistent in scoring than their native-speaker counterparts.

Caban (2003) examined the impact of raters' first language and educational training on oral assessments. Eighty-three raters were divided into four rater groups dependent on their L1 and academic rank: ESL-trained L1 English speakers, EFL trained L1 Japanese speakers, a group of ESL Japanese students, and English native speaker with no ESL background and no contact with Japanese speakers. The raters rated four videotapes of controlled oral interviews using seven rating categories: Grammar, Fluency, Content, Pronunciation, Pragmatics, Compensation Techniques, and Overall Intelligibility. Multifaceted Rasch Measurement was performed to investigate rating severities among groups. The findings revealed that there were similarities and differences among the four rater groups. For example, scores showed that all four rater groups judged Grammar, Compensation Techniques and Pronunciation with average difficulty and graded Content and Pragmatics with relative ease. Moreover, Fluency was found to be rated the most harshly while Overall Intelligibility the most leniently. Finally, the rater group that consisted of English native speaker with no ESL background and no contact with Japanese speakers was found to give the strictest scores while the other three rater groups were relatively lenient.

Kim (2009) examined how native English-speaking (NS) and non-native English-speaking (NNS) teachers assess students' oral English performance. Twelve Canadian NS teachers and 12 Korean NNS teachers were recruited as the two groups of raters. Ten Korean students' oral performance on three oral tasks was audio-recorded. The raters rated the speech samples according to a four-point rating scale. Written comments by the raters were provided along to justify the raters' ratings. The researchers used Many-faceted Rasch Measurement analysis to run quantitative data while having raters identify themes for evaluation criteria as qualitative analysis. The evaluation behaviors of the two groups were compared in terms of internal consistency, severity, and evaluation criteria. The results showed that most of the NS and NNS teachers maintained acceptable levels of internal consistency, and the two groups of raters exhibited similar severity patterns across different tasks. However, substantially dissimilar themes emerged in the evaluation criteria teachers used to justify students' performance. Nineteen themes were identified from qualitative data and then quantified and compared between the NS and NNS teacher groups. The total number of comments made by the two groups differed distinctly: while the NS group made 2,123 comments, the NNS group made 1,172, which demonstrated that NS teachers' judgments were more detailed and elaborate than those of the NNS teachers' judgments.

Xi and Mollaun (2009) (discussed in the accent familiarity section above) also set out to examine whether trained Indian raters were comparable with the U.S. operational raters in the context of the speaking section of the Test of English as a Foreign Language Test (TOEFL iBT). Results showed that with training, the Indian raters scored as consistently and accurately as did their counterparts in the U.S.

Zhang and Elder (2011) attempted to address the question of whether teachers who were native English speakers (NES) or non-native English speakers (NNES) rate differently in English oral unguided holistic assessment. Two groups of raters included 19 NES and 20 NNES teachers, who rated 30 speech samples elicited by the National College English Test-Spoken English Test (CET-SET) of China. Data were derived from two sources: unguided holistic ratings given by two groups of raters and the written comments that justify the ratings assigned. Quantitative data from MFRM analysis revealed no significant difference in raters' holistic judgments of the speech samples and demonstrated consistency between groups on the constructs of oral English proficiency. However, seven themes (Fluency; Content; Linguistic Resources; Interaction; Demeanor; Compensation Strategy; Other General Comments) identified from qualitative analysis of raters' comments showed different criteria in the way NES and NNES teachers graded. For example, NNES raters placed emphasis on Linguistic Resources as a main justification for their scores while NES raters focused on all the seven categories. This shows that NES utilized a wider range of evaluation criteria than NNES.

Baitman and Campos (2012) investigated the differences and similarities between native English speaker (NES) teachers and non-native English speaker (NNES) teachers in oral evaluations. Six NES raters and six NNES raters rated three speech samples based on the iBT TOEFL Test Independent Speaking Rubric and completed a questionnaire on their perceived importance of rated items. The results revealed that NES teachers were more lenient in the oral evaluation ratings than NNES teachers. NES teachers were also found to place more importance on fluency and pronunciation, while NNES teachers attended to grammatical accuracy and vocabulary.

In summary, most studies found no significant differences in terms of consistency and severity between the two groups of raters (Brown 1995; Caban, 2003; Kim, 2009; Zhang & Elder, 2011); however, the most recent study by Baitman and Campos (2012) found NES teachers tended to be more lenient than NNES teachers. To sum up, no consensus regarding the effect of first language background on rating behavior has yet been reached. In addition, all the studies report that the two groups of raters (NS and NNS) differed substantially on grading criteria. For instance, Kim (2009) found the NS group drew most frequently on overall language use, pronunciation, vocabulary, fluency, and specific grammar use whereas the NNS group emphasized pronunciation, vocabulary, intelligibility, overall language use, and coherence. Zhang and Elder (2011) showed NNES raters valued linguistic features while NES raters' comments were more widely distributed in the categories of Demeanor, Interaction, and Compensation Strategy. This may be partly due to the different backgrounds among the raters, marked differences among the contexts undertaken in the studies, particular features of the tests themselves as well as the nature of the evaluation scale used. As a result, different contexts might generate varying findings, thus more research is needed to provide in-depth comparison of NES and NNES raters under different contexts. Besides, these studies suffer from other limitations. First, the small and convenient (Zhang & Elder, 2011) sampling of participants make the results ungeneralizable to larger populations or applicable in different contexts. The second issue is that Kim (2009) and Zhang and Elder (2011) obtained data from holistic scores. Although the scorings revealed no difference between the raters, the written comments did demonstrate that the two groups of raters applied varying criteria on grading. Hence, there might have been fake consistency with points under different criteria offsetting holistic assessments. Future research should consider utilizing analytical scoring along with the holistic scoring to clarify and provide a more complete picture of the rating behavior between the two groups of raters. Another limitation is that the NNES raters used in most studies (Brown, 1995; Caban, 2003; Kim, 2009; Xi & Mollaun, 2009; Zhang & Elder, 2011) came from Asian educational background. Kim (2009) suggested that the instructional culture in EFL environments favor numerical scores and traditional fixed responses. This accordingly may have affected the way NNES raters formulated their comments in the studies. Therefore, different NNES raters, as well as analytical assessments are needed to further clarify the issues of offsetting points. Lastly, there seemed to be several variables within and across the chosen two groups of raters (Zhang & Elder, 2011). For instance, some teachers in the NNS groups had prior rating experience with the CET-SET test that was used in the study. Their former experience might unconsciously guide the way they rated in the study and thus clouded the results. In summary, the continuing debate on the status and value of native speaker norms and the uncertainty about how influential these are in different assessment contexts, give grounds for further exploration of the issue.

3.0 Summary of Research

In summary, although some studies found no significant differences between occupational practitioners and trained language raters (Huang, 2013; Lumley, 1998; Powers & Stansfield, 1985), some studies showed that one group of raters was harsher than the other (Barnwell, 1989; Brown, 1995; Hsieh, 2011). Secondly, while some studies demonstrated raters' accent familiarity with test takers' L1 could be a source of rater bias (Carey, Mannell & Dunn, 2011; Winke, Gass & Myford, 2012; Winke & Gass, 2013), some studies found the opposite (Huang, 2013; Xi & Mollaun, 2009). Lastly, most studies revealed that non-native speaker raters scored as consistently as native speaker raters (Brown 1995; Caban, 2003; Kim, 2009; Zhang & Elder, 2011); whereas one recent study (Baitman & Campos, 2012) found native-English-speaking teachers tended to be more lenient than their non-native counterparts. To conclude, no unanimity regarding raters' occupation, accent familiarity and native versus non-native speaker status on examiners' rating has been found so far.

4.0 Limitations & Future Research

In addition to the conflicting findings, all the studies, to some extent, shared similar weaknesses; namely sampling issues, different operational concepts, problematic designs, and lack of qualitative data.

The first limitation under the sampling issue is the small scale of the studies (Barnwell, 1989; Brown, 1995; Hsieh, 2011; Huang, 2013; Lumley1998; Powers & Stansfield, 1985; Zhang & Elder, 2011). In order to make the results more reliable and generalizable, larger-scale studies with more participants are therefore needed. The second problem is that the raters employed in the studies differed a lot in their backgrounds, which accordingly might have clouded the results (Hsieh, 2011; Zhang & Elder, 2011; Winke et al., 2012). To solve this problem, future research should recruit raters from similar backgrounds and as Winke et al., (2012, p.236) suggested, employ raters that better match the qualifications of operational raters in large-scale, language-testing-rating programs. Also, since each educational and testing context is unique (Kim, 2009), more research with varying contexts, different targeting examiners and differing measurements is needed.

Secondly, the studies held different operational concepts toward "accent familiarity" (Carey et al.,2011; Huang, 2013; Winke et al., 2012; Winke & Gass, 2013; Xi & Mollaun, 2009). Future research should be clear in the definition of accent familiarity and distinguish raters' accent familiarity with test takers' L1 stemming from L2 learning, bilingualism, residency in the country, or native speaker of the language. The results will therefore be more reliable and comparable among studies.

Additionally, there were flaws in the study design that weakened the studies (Hsieh, 2011; Huang, 2013; Kim, 2009; Zhang & Elder, 2011). For example, Huang's (2013) study would have been strengthened by having another group of raters, namely, non-familiar teachers. Future research should try to have a balanced study design. Another example is that some authors only looked at data derived from holistic scores (Kim, 2009; Zhang & Elder, 2011). That is, two speech samples may be rated the same for different reasons. One rater may place importance on pronunciation while the other rater may see content more importantly. Therefore, holistic scores are not only difficult for researchers to understand how a rater weights each item but also might mask the results. Future research could consider utilizing both holistic as well as analytical rating that help examining each rating criteria individually.

The last but not least limitation is that most studies relied on quantitative data in spite of having collected qualitative data like written comments (Baitman & Campos, 2012; Barnwell, 1989; Brown, 1995; Caban, 2003; Carey, Mannell & Dunn, 2011; Hsieh, 2011; Huang, 2013; Kim, 2009; Lumley1998; Powers & Stansfield, 1985; Winke, Gass & Myford, 2012; Xi & Mollaun, 2009; Zhang & Elder, 2011). The studies still did not provide information of how well each rater was actually rating. Future research should make use of qualitative data to substantiate the statistics by conducting interviews with raters to learn their process of assigning a grade to a particular speech sample, or interviews with biased raters to reveal the nature of rater bias.

5.0 Pedagogical Implications

Based on previous studies, the findings showed that raters' occupations, accent familiarity, and native language might potentially compromise the reliability of the test results, thereby affecting the decisions made in high-stakes exams. To prevent test scores from being affected by raters' characteristics rather than the examinees' actual performance, many researchers have called for improvements in rater training to tackle rater bias. However, some studies have shown that even with trained and certified raters, discrepancies still emerged among the raters (Ang-Aw & Goh, 2011; Barnwell, 1989; Carey, Mannell & Dunn, 2011; Hsieh, 2011; Lumley, 1998). Therefore, rater training should be reconsidered and regarded as an important part of raters' scoring. Moreover, an effective rater training package that contains specific modules should be designed to minimize the impact of irrelevant-constructs affecting raters' decisions. In addition to rater training, teachers could also help students to improve their speaking with some pronunciation techniques and simulated situational role-play practice. This literature review is intended to provide implications for raters as well as teachers and will discuss the possible procedures for the two professions respectively.

5.1 Implications for Raters

To start with, rater training is indispensable in testing programs and inseparable from the testing itself. In the absence of rater training (Caban, 2003; Kim, 2009; Zhang & Elder, 2011), it may not be surprising that raters' different backgrounds guided and determined their rating process (Brown, 1995). In addition to the outcome of the rating, the raters in Xi & Mollaun's (2009) study reported feeling more confident in scoring after the training. Hence, rater training is needed to help raters with different backgrounds recalibrating their standards so that all rate on the same page and also to increase raters' confidence in grading.

Secondly, as Ang-Aw and Goh (2011) suggested, rater training should talk about the aspects of candidates' performances upon which the raters should and should not focus. That is, the training should clearly define the criteria represented in the rating scale. To do that, the training could have raters practice identifying the characteristics of candidates' performances that correspond to the descriptors in the marking scheme (Pollitt & Murray, 1996). By practicing identifying several samples, trainees could have a better sense of what is important from what is not on the assessment scale.

Next, rater training programs should address how raters' linguistic experience (raters' L1 or familiarity with other accents) may lead to bias in assigning grades. The drill could have raters practice listening and rating speech of a variety of language groups. The goal of this would be to familiarize trainees with different test takers' accents and by extension to test takers' nonnative-like word- and sentence-level processing (Bradlow & Bent, 2008). This type of training, as recommended by Carey et al. (2011) and Xi and Mollaun (2009), could also help trainees become aware of how one's linguistic experience might inadvertently influence their ratings.

Additionally, rater training should include instruction on differentiating difficulties in comprehension that stem from pronunciation errors from cognitive loads in attending to unfamiliarly accented speech. This cognitive demand occurs when the speech samples comprise different foreign accents. Take the raters in Winke and Gass 's (2013) study for an example. The raters reported in their written comments that their lack of awareness of what the next foreign accent was increased their cognitive load, thereby affecting their rating process. On the other hand, the research in which raters graded a single type of foreign accent (e.g., English-accented Japanese in Brown, 1995; Korean-accented English in Kim, 2009; Chinese-accented English in Zhang & Elder, 2011) resulted in no differences in consistency or severity among raters grouped by their linguistic background because there was no accent variation among the speech samples. Therefore, language testing programs that have test takers from various L1 backgrounds need to address this issue. Another way to deal with this problem is that they could sort out the speech samples by L1 background for raters. That way, raters could consistently and reliably rate the samples of speech produced by test takers of that L1 and then proceed to rate speech samples of other languages.

Wigglesworth (1993) found that providing raters with feedback on their rating behavior between oral rating sessions made them more consistent in subsequent sessions. To elaborate, eight raters in Wigglesworth's (1993) study were recalled after their trial rating for a refresher rater-training session. During the session, each rater was provided with individualized feedback concerning his or her earlier trial rating with respect to the analyses determined by the bias system. The goal of this session was to provide raters with an 'assessment map' showing whether the raters had behaved unusually on any item of the rating scale. The results of the study proved the effectiveness of rater feedback and again indicated that giving qualitative feedback to raters during training helps reduce rater bias.

Another issue is that native- and non-native English speaking raters applied different criteria when scoring (Kim, 2009; Zhang & Elder, 2011). For example, a native speaking rater would be more likely to pick up and comment on features of interaction, whereas a non-native speaker would be more likely to focus on linguistic resources (Zhang & Elder, 2011). Hence, for the two groups of raters to have more consistency in scoring, language testing programs that have native and non-native speaker raters could include sessions in which both groups of raters reflect and discuss about their criteria and judgments applied in rating.

What is also important is that the training should lead raters to reflect and examine their own rating process. By doing so, raters could first observe 'expert raters' or 'experienced raters' decision-making process (Koh, 2003). For instance, the 'expert raters' could demonstrate their rating process by using think-aloud method. That way, the trainees could follow an expert rater’s thoughts and learn through the decision-making process. The goal of this training would be to help raters reflect more critically upon the reasons for a particular speech sample and the steps involved in arriving at a rating decision.

As effective and promising as the impact of the rating training seems; however, it could not solve all the problems stemming from rater characteristics. Even trained raters could produce a lot of variability in scoring. To illustrate, Ang-Aw and Goh (2011) aimed to find whether there were any differences in experienced rater judgments in a high-stake oral examination in Singapore. Seven trained and experienced raters of the ‘O’ Level oral examination were selected and rated four students’ speech samples from the test. Data were derived from Concurrent Verbal Protocols (CVP), a questionnaire and scores. The results revealed that although the raters underwent similar training and were considered experienced raters, differences still existed in four areas: emphases on criteria, perceived constructs of oral proficiency, interpretations of performances and scores, and approaches in assessment. Hence, as Ang-Aw and Goh (2011) suggested, even with trained raters, testing raters’ inter-rater reliability before their actual rating is still crucial.

Lastly, according to Lumley and McNamara (1995), it is impossible to completely eliminate rater variability even through rater training. In their study, they attempted to address the question of stability of rater characteristics and rater bias over two years in the context of the speaking subtest of the Occupational English Test (OET). Eleven speech samples were derived from participants’ two role-play based interviews. Data were presented to 13 raters at two rater training sessions separated by an 18-month interval and a follow-up operational test administration session. The study found significant changes in rater severity through time while the extent of rater bias was not large. As a result, the outcome of rater training may not endure for long after a training session. Consequently, as Lumley and McNamara (1995) recommended, holding a moderation session before each test administration is needed, which will allow raters to re-orient an internalized set of criteria for their ratings.

5.2 Implications for Instructors

First, the finding that listeners adapt to a certain accented speech when repeatedly presented with it offers two implications for teachers. To begin with, language teachers could take advantage of the finding and train students’ listening skills by providing students with abundant auditory materials. According to Krashen’s Input Hypothesis, a large amount of input favors learners’ language learning. It is especially important in an EFL environment where English is only spoken in the classroom setting. Therefore, by presenting students with listening materials produced by native-speakers, learners could adapt to the accent thereby understanding native speakers better as well as have a model to imitate. Additionally, by presenting students with various accented speech samples, learners could understand the concept of accent and are likely to comprehend others who speak the target language with a different accent. Another way to look at the statement is that language teachers may have become better listeners of their students' speech because of teachers' contact with the students. Thus, language professionals could use their keen ears in identifying individual student's speech case and point out the problems explicitly to the students. Then as Huang (2013) suggested, teachers need to pay more attention to pronunciation and speech instruction, particularly for students with strong accents, to help them not only score better on high-stakes exams but also achieve effective communication.

Another implication is regarding the international teaching assistants (ITAs). At large research universities, nearly all undergraduate students report that they have experience with ITAs over a course (Williams, 2006). However, it is sometimes reported that the ITA is a problematic character in universities (Fitch & Morgan, 2003) because students often complain about ITAs' lack of speaking clarity, thereby tending to downgrade their speech (Fox & Gay, 1994). Accordingly, in order to help ITAs become more intelligible for undergraduate students, pronunciation classes should be designed to assist in improving ITAs' speaking skills. To illustrate, Sardegna (2012) found that after teaching international graduate students how to use a variety of English pronunciation strategies, and having the students practice integrating the strategies when they spoke, these strategies not only helped improve students' pronunciation but also equipped the students with the ability to self-correct and self-teach. Therefore, English pronunciation lessons that provide strategies to deal with individual sound problems, rhythm, thought groups, primary stress, linking, and intonation (Sardegna, 2010) should be given to help ITAs. Through raising ITAs' awareness of English pronunciation rules explicitly, ITAs could become more intelligible for undergraduate students; and undergraduate students will then have more positive experiences with ITAs.

The last implication concerns the use of the target language in occupational contexts. To elaborate, many occupational tests regard both examinees' language ability as well as the capacity to cope with the work as two crucial criteria for rating. As a result, occupational examinations conduct a series of interviews that stimulate some aspects of the role which the examinees will be playing in their professional fields. For example, Brown (1995) utilized several face-to-face situational role-plays in her study of the Japanese tourist guide, such as testing examinees how to help a client to choose a tour, how to deal with an upset or worried client, how to present cultural aspects of a topic in an interesting way, and how to clearly explain an itinerary to clients. Thus, language teachers should not only facilitate students' language learning but also foster students' ability that will help learners to handle difficult situations using appropriate language in their professional contexts.

6.0 Conclusion

Language testing researchers have been studying the impact of rater effects on raters' scorings for a long time. That is, in order to make test results impartial and reliable, researchers strive to find potential inconsistencies introduced into the rating process by the raters themselves. This literature review examined the research that explored rater effects and discussed the rater variability on language oral evaluations in three respects: raters' occupations, raters' accent familiarity and raters' native language. In summary, while some studies found no significant differences between vocational practitioners and trained language raters (Huang, 2013; Lumley, 1998; Powers & Stansfield, 1985), other studies revealed that one group of rater was stricter than the other (Barnwell, 1989; Brown, 1995; Hsieh, 2011). Next, some studies demonstrated rater bias from raters' accent familiarity with test takers' L1 (Carey, Mannell & Dunn, 2011; Winke, Gass & Myford, 2012; Winke & Gass, 2013); yet other studies did not find a significant relation between rater bias and raters' accent familiarity (Huang, 2013; Xi & Mollaun, 2009). Finally, most studies revealed that non-native speaker raters ' ratings were as consistent as that of native speaker raters' (Brown 1995; Caban, 2003; Kim, 2009; Zhang & Elder, 2011); whereas a recent study (Baitman & Campos, 2012) indicated that native-English-speaking teachers were more lenient than their non-native counterparts. In conclusion, no agreement concerning raters' professions, accent familiarity and native-speakerness v. non-native-speaker status on examiners’ rating has been found. In short, the impact of rater characteristics on oral assessments remains unclear. Moreover, despite the three rater characteristics discussed in this paper, other raters’ characteristics such as raters’ scoring experience (Cumming, 1990; Weigle, 1994, 1999), trained or untrained (Ang-Aw & Goh, 2011; Barnwell, 1989; Hsieh, 2011; Huang, 2013), and gender (Liu, 2011) are also given rise to future research.

In conclusion, in order to have fair and reliable testing scores, rater characteristics leading to rater bias calls for further research to unveil the effect of rater characteristics as well as to provide a deeper insight of the nature of rater biases.

References

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

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, *12*, 238-257.

Baitman, B., & Campos, M. V. (2013). A comparison of oral evaluation ratings by native English speaker teachers and non-native English speaker teachers. *Literatura y Lingüística*, *27*, 171-200.

Barnwell, D. (1989). 'Naïve' native speakers and judgments of oral proficiency in Spanish. *Language Testing, 6*, 152–163.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*, 707–729.

Brown, A., Iwashita, N., McNamara, T., 2005. An Examination of Rater Orientations and Test-taker Performance on English-for-Academic-Purposes Speaking Tasks. *Research Report-Educational Testing Service, Princeton RR*, 5.

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

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese

ESL students. *Second Language Studies*, *21*, 1-43.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*, 201–219.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*, 3647-3658.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando, FL: Holt, Rinehart andWinston.

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

Davies, A. (1999). Standard English: Discordant voices. *World Englishes, 18*, 171–186.

Davies, A. (2003). *The native speaker: Myth and reality* (2nd ed.). Buffalo, NY: Multilingual Matters.

Fitch, F., & Morgan, S. E. (2003). "Not a lick of English": Constructing the ITA identity through student narratives. *Communication Education*, *52*, 297–310.

Fox, W., & Gay, G. (1994). Functions and effects of international teaching assistants. *Review of Higher Education*, *18*, 1–24.

Hsieh, C. N. (2011). Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spain Fellow Working Papers in Second or Foreign Language Assessment 9*, 47-74.

Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *System*, *41*, 770-785.

Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press.

Liu, W. (2011). Gender and occupational knowledge: the effects of rater variability on the assessment of ESP oral performance. (Unpublished)

International Language Testing Association. (2005). Code of ethics for ILTA. Retrieved from http://www.iltaonline.com/code.pdf

Joseph JE (2004). *Language and identity: National, ethnic, religious*. Basingstoke, UK: Palgrave Macmillan.

Kang, O. (2012). Impact of Rater Characteristics and Prosodic Features of Speaker Accentedness on Ratings of International Teaching Assistants' Oral Performance. *Language Assessment Quarterly*, *9*, 249-269.

Kim, Y. (2009). An investigation into native and non-native teachers’ judgments of oral English performance: A mixed methods approach. *Language Testing, 26*, 187–217.

Koh, C. C. H. (2003). An exploratory study of three raters' decision-making process on the picture conversation task used for primary six candidates in Singapore.

Kunnan, A. J. (2000). Fairness and justice for all. *Fairness and Validation in Language Assessment*, *9*, 1-14.

Lowenberg, P. H. (2002). Assessing English proficiency in the Expanding Circle. *World Englishes, 21*, 431–435.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*, 238–257.

Luoma, S. (2004). *Assessing speaking*. Ernst Klett Sprachen.

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, *36*, 173-190.

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

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28*, 111–131.

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

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. *Language Testing Research Colloquium (LTRC), Cambridge and Arnhem, 3*, 74-91.

Powers, D. E., & Stansfield, C. W. (1985). Testing the oral English proficiency of foreign nursing graduates. *The ESP Journal*, *4*, 21-35.

Sardegna, V. G. (2010). Pronunciation learning strategies that improve ESL learners' linking. *Pronunciation and Intelligibility: Issue in Research and Practice, 104*

Sardegna, V. G. (2012). Learner differences in strategy use, self-efficacy beliefs, and pronunciation improvement. (Unpublished)

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

Stanlaw, J. (2004). Japanese English: Language and culture contact. Hong Kong: Hong Kong University Press.

Taylor, L. B. (2006). The changing landscape of English: Implications for language assessment, *ELT Journal, 60*, 51–60.

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

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*, 305-319.

Williams, G. (2006). Cultural, professional and personal influences on the teaching identity development of international teaching assistants (Unpublished doctoral dissertation). University of Georgia, Athens.

Winke, P., & Gass, S. (2013). The Influence of Second Language Experience and Accent Familiarity on Oral Proficiency Rating: A Qualitative Investigation.

TESOL Quarterly, 47, 762-789

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

Xi, X., & Mollaun, P. (2009). How do raters from India perform in scoring the TOEFL iBT*TM* speaking section and what kind of training helps? (TOEFL iBT Research Report RR-09–31). Princeton, NJ: Educational Testing Service.

Yano Y (2001). World Englishes in 2000 and beyond. *World Englishes, 20*, 119–132.

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