

TOEFL iBT® Research Report
TOEFL iBT-16

**The Relationship Between Raters’
Prior Language Study and the
Evaluation of Foreign Language
Speech Samples**

Paula Winke

Susan Gass

Carol Myford

July 2011

**The Relationship Between Raters' Prior Language Study and the Evaluation of Foreign
Language Speech Samples**

Paula Winke and Susan Gass
Michigan State University

Carol Myford
University of Illinois at Chicago



ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

Copyright © 2011 by ETS. All rights reserved.

No part of this report may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher. Violators will be prosecuted in accordance with both U.S. and international copyright laws.

ETS, the ETS logos, GRADUATE RECORD EXAMINATIONS, GRE, LISTENING, LEARNING. LEADING., TOEFL, TOEFL IBT, and the TOEFL logo are registered trademarks of Educational Testing Service (ETS).

COLLEGE BOARD is a registered trademark of the College Entrance Examination Board.

Abstract

This study investigated whether raters' second language (L2) background and the first language (L1) of test takers taking the TOEFL iBT[®] Speaking test were related through scoring. After an initial 4-hour training period, a group of 107 raters (mostly of learners of Chinese, Korean, and Spanish), listened to a selection of 432 speech samples that 72 test takers (native speakers of Chinese, Korean, and Spanish) produced. We analyzed the rating data using a multifaceted Rasch measurement approach to uncover potential biases in the rating process. In addition, 26 of the raters participated in stimulated recall sessions, during which they watched videos of themselves rating. Using the video as a prompt, we asked them to discuss and explain their rating processes at the time of rating. The results from our bias interaction analyses revealed that matches between the raters' L2 and the test takers' L1 resulted in some of the raters assigning ratings that were significantly higher than expected. As a whole, raters with Spanish as an L2 were significantly more lenient toward test takers who had Spanish as an L1, and raters with Chinese as an L2 were significantly more lenient toward test takers who had Chinese as an L1. Analyses of the qualitative data, assisted by the program QSR NVivo 8, revealed information concerning the raters' awareness of their biases.

Key words: oral assessment, second language performance assessment, item response theory (IRT), rater performance, rater bias, Rasch measurement, Facets, NVivo

TOEFL[®] was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board[®] assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations[®] (GRE[®]) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education. The test is now wholly owned and operated by ETS.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, two-year colleges, and nonprofit educational exchange agencies.



Since its inception in 1963, the TOEFL has evolved from a paper-based test to a computer-based test and, in 2005, to an Internet-based test, TOEFL iBT[®]. One constant throughout this evolution has been a continuing program of research related to the TOEFL test. From 1977 to 2005, nearly 100 research and technical reports on the early versions of TOEFL were published. In 1997, a monograph series that laid the groundwork for the development of TOEFL iBT was launched. With the release of TOEFL iBT, a TOEFL iBT report series has been introduced.

Currently this research is carried out in consultation with the TOEFL Committee of Examiners. Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, solicits, reviews, and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Current (2010-2011) members of the TOEFL Committee of Examiners are:

Alister Cumming (Chair)	University of Toronto
Carol A. Chapelle	Iowa State University
Barbara Hoekje	Drexel University
Ari Huhta	University of Jyväskylä, Finland
John M. Norris	University of Hawaii at Manoa
James Purpura	Columbia University
Carsten Roever	University of Melbourne
Steve Ross	University of Maryland
Mikyuki Sasaki	Nagoya Gakuin University
Norbert Schmitt	University of Nottingham
Robert Schoonen	University of Amsterdam
Ling Shi	University of British Columbia

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Acknowledgments

This work was funded by the TOEFL[®] Research Program at Educational Testing Services (ETS). The authors would like to thank Dan Reed, Dennie Hoopingarner, Megan Kilbourn, four anonymous reviewers from ETS, and the ETS TOEFL Research Committee for their generous support and assistance.

Table of Contents

	Page
Executive Summary	1
Overview	2
Literature Review.....	4
Rater Effects in Performance Assessment.....	4
Rater Effects in Writing Assessment.....	5
Rater Effects in Speaking Assessment	7
Training as a Tool to Mitigate Rater Effects in Assessment	8
Accent Familiarity and Rating.....	8
The Need for Further Research Into Rater Effects in Speaking Assessment	13
Using Introspection to Study Rater Effects in Performance Assessment.....	14
Research Question	15
Methodology.....	16
Participants	16
Materials	16
Data Collection Design.....	18
Procedure	19
Data Analysis.....	21
Results.....	27
Quantitative Results.....	28
Qualitative Findings.....	41
Discussion.....	47
Conclusion	50
Next Steps.....	53
References.....	54
Notes	62
Appendices	
A - Background Questionnaire.....	63
B - Sample Rating Sheet	66

List of Tables

	Page
Table 1. Rater Background Characteristics.....	17
Table 2. Distribution of the 432 Sound Files by L1, Gender, and TOEFL iBT Test Score	18
Table 3. L2 Background and Gender of the Stimulated Recall Participants.....	19
Table 4. Item Facet Summary Statistics (Sorted by Infit Mean-Square Values)	26
Table 5. Differences in Severity/Leniency That Selected MSU Raters Exhibited	27
Table 6. Test Taker L1 Subgroup Measurement Report.....	34
Table 7. Differences in L1 Test Taker Subgroups' Average Proficiency Measures.....	36
Table 8. Bias/Interaction Report for Test Taker L1 Subgroups and Rater L2 Subgroups.....	38
Table 9. Rater L2 Subgroup Measurement Report	40
Table 10. Rater ESL/EFL Teaching Experience Subgroup Measurement Report.....	40
Table 11. Summary of Coded Data From Stimulated Recalls	42

List of Figures

	Page
Figure 1. The study's data collection design.	21
Figure 2. Variable map from the FACETS analysis of the data.	30
Figure 3. Bias/interaction analysis specified by test taker L1 and rater L2.	39

Executive Summary

Testing programs want to keep rater background characteristics from inappropriately influencing the rating of speech samples. Thus, one of the main tasks of testing programs is to identify which rater background characteristics influence scores. After identifying characteristics that have an effect on scores, rater training programs can be developed to mitigate the effects of those characteristics on scoring.

As of yet, no researchers have investigated whether a rater's prior study of a second language (L2) that matches the first language (L1) of a test taker influences the rater's rating of that test taker. Thus, the present study seeks to determine to what extent a rater's knowledge of the test taker's L1, even when the rater's knowledge of the test taker's L1 is non-native-like, may influence the rater's evaluation of the test taker's recorded L2 speech. In some cases, a rater may be aware of such an influence. In other cases, the rater might be totally unaware that knowledge of the test taker's L1 exerts an influence. Therefore, this study also sought, through qualitative data analysis, to discover whether raters are aware of such biases, if they exist. More specifically, the research question that guided this study was the following: Are there certain groups of trained raters (grouped by their L2) who exercise differential severity, depending on the L1 of the test taker?

After an initial 4-hour training period, a group of 107 raters, who were mostly undergraduate learners of Chinese ($n = 41$), Korean ($n = 11$), and Spanish ($n = 48$), listened to a selection of 432 speech samples that 72 test takers (24 native speakers of Chinese, 24 of Korean, and 24 of Spanish) produced. We analyzed the rating data using a multifaceted Rasch measurement approach to uncover potential biases in the rating process. In addition, 26 of the raters participated in stimulated recall sessions, during which they watched videos of themselves rating. Using the video as a prompt, we asked them to discuss and explain their rating processes at the time of rating.

Results from the bias interaction analysis indicated that the raters with Spanish as an L2 were significantly more lenient toward the Spanish-native-speaking test takers than were the raters with either Korean or Chinese as an L2. The raters with Chinese as an L2 were significantly more lenient toward the Chinese-native-speaking test takers than the other raters were. Analyses of the qualitative data, assisted by the program QSR NVivo 8, revealed that some raters' prior L2 learning experiences may have interacted with the L1s of the test takers. A

number of the raters commented that they often thought about the test takers' L1 accents, and that they frequently tried to identify the test takers' L1s. They acknowledged that they were most successful in doing so when their own L2 was the L1 of the test takers. In many cases, it appeared that the test taker's accent seemed to affect the way these raters perceived the quality of the test taker's speech. Some of the raters who were heritage language learners appeared to sense the impact of the test takers' L1 on their own rating behavior. Their interactions with the test takers' L1 at times brought out strong emotive aspects and made them think about their own language learning processes and linguistic backgrounds. These interactions seemed to be a natural part of the raters' listening processes, regardless of whether they affected the assignment of ratings.

The qualitative findings together with the bias analysis results lead us to suggest that when raters know, to varying degrees, the L1 of the test takers and can discern the L1 of the test takers through the test takers' ethnic accents, that knowledge may influence their ratings. Whether these language-background-related biases would be found to exist within populations of raters who are highly trained, certified, and experienced needs further investigation.

Overview

The assessment of speaking ability is challenging, especially when carried out on a large scale and in a standardized testing context. One of the many reasons why assessing speaking is difficult may be its socio-affective nature. There are factors that influence people's impression of what it is that qualifies speech as being good or fluent. Further, these ideas may be reflected in the ratings of speech samples, even when test designers attempt to lessen the impact of these influences through the design of a thorough and detailed scoring rubric, through intensive rater training, and even through rater retraining efforts. To further limit subjectivity in rating and to maximize the reliability of ratings, high-stakes testing programs often require raters to meet a number of background requirements, complete a rater training program, and pass a qualifying test before they can be certified to rate.

However, while testing programs want to keep rater background characteristics from inappropriately influencing the rating of speech samples, they are limited in the amount of time, space, and money available to try to ensure that background characteristics do not influence ratings. Testing programs need rater training that is cost effective and efficient. While detailed scoring rubrics may be ideal for properly directing the rating process, for ease in use, the scoring

rubrics need to be easy to apply. Thus, questions arise as to how short the training can be, and how detailed the scoring rubric should be. Likewise, appropriately large pools of qualified and reliable raters are difficult to assemble. Testing programs want to restrict diversity in the rating pool to ensure consistency in rating, yet they also want to be able to cast a wide enough net to make certain that they have a sufficient supply of capable raters. Testing programs therefore need to know which rater background characteristics are likely to influence ratings, and which ones aren't, so that they may appropriately increase the pool of potential raters without compromising assessment quality.

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education, 1999) emphasize the importance of building a scientifically sound and convincing validity argument in order to support the validity of inferences that test users make from test scores. For assessments that involve raters who are applying criteria to judge test takers' performance, the *Standards* underscore the need to determine "whether [the raters] are, in fact, applying the appropriate criteria and not being influenced by factors that are irrelevant to the intended interpretation" (p. 13). To gather such evidence, researchers are instructed to conduct "empirical studies of how observers or judges record and evaluate data along with analyses of the appropriateness of these processes to the intended interpretation or construct definition" (p. 13). During a rating procedure, if raters are influenced by factors that are not included in the scoring rubric that they are supposed to be applying, then those factors may be introducing unwanted construct-irrelevant variance into the rating process, thus clouding the interpretation of the scores from the assessment. Therefore, as part of the validity evidence to support inferences from scores on standardized speaking tests, studies are needed to determine whether raters evaluating test takers' speech samples are influenced by any extraneous factors (such as rater or test taker background characteristics) that may impact their ratings of test taker performance and thus distort the meaning of the scaled scores and the inferences that test users might make from those scores.

In fact, one could argue that this type of empirical research is ethically mandated, at least in an indirect fashion. The most recent publication on ethics for language testers, compiled by the International Language Testing Association (2005), states, "language testers shall not discriminate against nor exploit their test takers" on the grounds of the test takers' background

criteria, and that language testers “should not impose their own values, to the extent that they are aware of them” (p. 2). What is important is that, unbeknownst to the test designers and even to the raters themselves, raters may impose their own values. Studying the influences of rater background characteristics on raters’ ratings is thus crucial to controlling for sources of potential bias in a rating procedure. It is vital to gain an understanding of what rater and/or test taker background characteristics might influence raters during the rating process. In other words, it is important to look at the interactions between rater and test taker to see if there is any evidence of scoring biases in those interactions.

Normally native or near-native speakers of the targeted language rate speech samples. But these raters often are bilingual or multilingual, having studied one or more languages that may happen to be the native languages of the test takers they rate. Testing programs rarely report on the backgrounds and experiences of their individual raters outside of their shared L1. Rather, they frame the group of raters as a homogeneous entity. Few researchers have conducted studies to determine the extent to which ratings that individual raters assign may be influenced by their multilingualism or prior language learning experience. Researchers have thus far only investigated bilingual raters who are near-native speakers of the target language being tested (e.g., Johnson & Lim, 2009; Xi & Mollaun, 2009). They have sought to understand whether the bilingual raters’ and test takers’ shared L1 influences the bilingual raters’ assignments of ratings. No researchers have investigated whether raters’ prior study of an L2 that happens to match the L1 of test takers influences their assignment of ratings. Thus, the present study adds to the current literature by seeking to determine to what extent a rater’s knowledge of the test taker’s L1, even when the rater’s knowledge of the test taker’s L1 is non-native-like, may influence the rater’s evaluation of the test taker’s recorded L2 speech. In some cases, a rater may be aware of such an influence. In other cases, the rater might be totally unaware that knowledge of the test taker’s L1 exerts an influence. Therefore, this study also seeks, through qualitative data analysis, to discover whether raters are aware of such biases, if they exist.

Literature Review

Rater Effects in Performance Assessment

The study of rater effects in performance assessment has a long history in language testing. Lumley and McNamara (1995) noted that, for at least a century, researchers have investigated variability in rater performance in a variety of language testing contexts. Using

multifaceted Rasch measurement (Andrich, 1988; Bond & Fox, 2001; Linacre, 1989) as implemented in the computer program FACETS (Linacre, 2009), researchers have gained an understanding of how raters differ in the levels of severity they exercise when rating various groups of test takers (Lumley & McNamara, 1995). (For a detailed review of L2 testing research that has used Rasch measurement to estimate patterns of rater bias, see Schaefer, 2008.)

Studies that have addressed rater effects in performance assessments can be divided into two categories: (a) those that have investigated rater effects within the context of writing assessment, and (b) those that have investigated rater effects within the context of speaking assessment. This categorization is necessary because writing assessment and oral assessment differ in important qualitative ways. Raters of writing performance do not have to contend with linguistic features such as pronunciation, hesitations, intonations, and pragmatics that exist in the oral context (Johnson & Lim, 2009)—features that may interact with raters as listeners. Our study is concerned with oral assessment. Nonetheless, before reviewing rater bias research in the context of oral assessment, we provide a short and condensed review of studies of rater bias in the context of writing assessment, because rater bias studies in the oral context employ many of the same methodologies, and, in some cases, the outcomes are similar.

Rater Effects in Writing Assessment

In a review of 70 studies concerning ESL/EFL essay tests, Barkaoui (2007) found that 22 examined rater effects. Barkaoui reported that rater characteristics such as personality; cultural, linguistic, and educational background; and teaching and rating experience influence raters' decision-making processes, behaviors, and levels of severity and self-consistency. Barkaoui found that the rater characteristics that received the most attention in the writing literature were raters' L1 background, academic background, and prior teaching and rating experience. What concerns us most are studies on raters and L1 background. These have shown mixed results, with some studies reporting no differences in ratings that native speakers and nonnative speakers of the language being assessed have assigned (Connor-Linton, 1995; Johnson & Lim, 2009; Shi, 2001), while other studies have reported differences (Hamp-Lyons, 1989; Hamp-Lyons & Zhang, 2001; Hill, 1997; Kobayashi, 1992). Below, we look at two of these studies, one from each group, to understand how they arrived at differing conclusions.

Kobayashi investigated how English native speakers and Japanese native speakers differed in their evaluations of English compositions that university-level Japanese students

wrote. One hundred and forty-five English-native-speaking raters and 124 Japanese-native-speaking raters used a 10-point scale to evaluate two compositions on grammar, clarity of meaning, naturalness, and organization. The relevant finding for Kobayashi's study was that the English native speakers were found to be harsher in terms of rating grammaticality than the Japanese native speakers. However, it must be noted that Kobayashi did not measure the Japanese native speakers' English language proficiency. One might conclude that their non-native-like abilities in English prevented them from adequately evaluating errors in grammaticality. It is unclear whether they were more lenient in evaluating grammaticality because they themselves were unsure of how to evaluate grammar, or because they were nonnative speakers who sympathized with the test takers. Hamp-Lyons (1989), Hamp-Lyons and Zhang (2001), and Hill (1997) all reported similar findings. Raters who shared an L1 with the essay writers tended to be more lenient in assigning ratings. But in each case, the implications for high-stakes testing programs are not clear because such programs do not normally employ raters who are non-native-like in the language being tested. In all fairness to these studies' authors, the goal of these studies was not to determine whether non-native-like speakers of a writing test's target language should be raters in high-stakes testing programs; rather, the studies simply pointed out that raters with different language backgrounds and proficiencies in the target language valued and used rating criteria differently. This is not surprising, especially because the raters were untrained—they were left to their own devices to use and properly interpret the categories on the scoring rubrics and rating scales.

Johnson and Lim (2009) found that it is possible to train native-like, nonnative speakers of English to rate English compositions as effectively as their native-English-speaking rater counterparts. They investigated 17 official Michigan English Language Assessment Battery (MELAB) raters from the English Language Institute at the University of Michigan. These raters included four bilingual speakers—two of Spanish and English, one of Korean and English, and one who had English, Amoy, and Tagalog as L1s. Using a multifaceted Rasch measurement approach, they analyzed the ratings these raters gave to 7,400 examinees over 3 years. The results showed that the Korean-English bilingual speaker exhibited a slight bias—she was somewhat more lenient when rating compositions that Korean L1 examinees wrote. The researchers found evidence of other biases related to language background in the data for both types of raters (native speakers and nonnative speakers); but, as Johnson and Lim pointed out,

the small number of significant bias interaction terms spread across the raters made it difficult to interpret the findings as showing differences between native and near-native speakers in their patterns of bias. Noting that for the MELAB scoring, two raters are always used, they concluded that their results showed that the four native-like, nonnative speakers were just as accurate and consistent in rating writing performance as their native-speaking peers.

Rater Effects in Speaking Assessment

Similar to researchers studying writing assessments, researchers studying speaking assessments have paid relatively little attention to the L1 background of the raters and the differences in ratings that native speakers and nonnative speakers of the language being assessed have assigned (Brown, 1995; Fayer & Krasinski, 1987; Kim, 2009; Xi & Mollaun, 2009). As Kim (2009) and Xi and Mollaun (2009) explained, one reason why native and nonnative speakers of the language being assessed are currently examined in the oral assessment context is because administrators in charge of these testing programs want to know whether nonnative speakers can appropriately use a scoring rubric to evaluate oral speaking proficiency. Some researchers have reported that in oral assessment contexts, raters from diverse L1 backgrounds tend to use the categories on a scoring rubric in a different manner than their native-speaking counterparts (Brown, 1995). However, the results from these studies are complicated and mixed, with researchers reporting contradictory findings. For example, some researchers have found that nonnative raters were more severe than native raters (Fayer & Krasinski, 1987); by contrast, other researchers have found that native speakers were more severe (Hill, 1997), yet not significantly so (Brown, 1995). Still other researchers have shown that quantitatively there were no differences in rater severity between native and nonnative speakers (Kim, 2009; Xi & Mollaun, 2009).

Comparing these studies is problematic due to differing methodologies. Fayer and Krasinski (1987) used untrained raters who did not employ a detailed scoring rubric and only rated one speech task per test taker. Brown (1995) and Kim (2009) employed raters who varied not only in their L1 backgrounds but also in their years of teaching experience, with some raters having no teaching experience, and others having a great deal of experience. And Xi and Mollaun's (2009) study included 26 trained raters who were bilingual speakers of an Indian dialect and English, who could arguably be considered native speakers of English. Thus, a large-scale study is needed that will investigate differences in ratings assigned by those who know, and

those who do not know, the L1s of the speakers being tested. Such a study should control for teaching experience and use trained raters.

Training as a Tool to Mitigate Rater Effects in Assessment

As Xi and Mollaun (2009) noted, a reason why administrators of assessment programs want to identify rater biases is so that they can then create rater training programs that will explicitly address and attempt to minimize the influences of known biases. High quality rater training can help raters better understand the categories and criteria represented in the rating scale (Saito, 2008; Weigle, 1994, 1998), which may affect their rating behavior. Wigglesworth (1993, 1994) found that providing raters with feedback on their rating behavior between oral rating sessions (i.e., charts of their patterns of biases, estimated through multifaceted Rasch analysis) made them more consistent in subsequent oral rating sessions. Kondo-Brown (2002) also reported that training between essay rating sessions improved raters' internal consistency. Saito (2008) found that longer rater training did not improve peer ratings of oral performance, but it had an effect on rating behavior—in particular, it may have helped raters adopt a common frame of reference. Rinnert and Kobayashi (2001) argued that raters in their study showed that they could change their rating behavior even without explicit rater training—they observed that as native-Japanese-speaking English essays raters gained more experience in teaching English and rating English essays, they moved from preferring writing that contains features of their L1 (Japanese) to preferring writing that contains features of their L2 (English). The results from Rinnert and Kobayashi's study were similar to those from Cumming (1990) and Wolfe, Kao, and Ranney (1998), who found that expert or proficient raters of writing, when compared to novice or less proficient raters, were more likely to be accurate in judging written language samples. Thus, participating in rater training, becoming more experienced in teaching, or gaining experience in rating English essays can lead to better (more accurate and more consistent) rater performance. On the other hand, Lumley and McNamara (1995) noted that it is impossible to completely eliminate rater variability, even through training. But training may help minimize any observable biases in ratings.

Accent Familiarity and Rating

By investigating interactions that may occur in an oral testing context when the raters' L2 and the test takers' L1 are shared, we sought to determine the effects of accent familiarity on

raters' rating processes. When a person learns a second or foreign language after childhood, it is natural that when the person speaks the foreign language, there will be an accent (Major, 2001; Munro, Derwing, & Morton, 2006). When listeners hear the nonnative speaker speak, they are extremely adept at identifying the presence of the foreign accent (Flege, 1984), that is, noticing that the speech differs in quality from their own (Derwing & Munro, 2005). Research has shown that listeners can identify a foreign accent easily, even when the speech is a single word utterance played backwards (Munro, Derwing, & Burgess, 2003).

Although it is very interesting to speech perception researchers that listeners can identify a speaker as a nonnative speaker extremely well, for language testers, what is more interesting is how the listener reacts to that piece of information. The listener may identify the speaker as nonnative (Esling & Wong, 1983); subsequently, in real-life face-to-face conversations, the social interactions between the nonnative speaker and listener may be altered (Derwing & Munro, 2005). For example, the listener may consequently provide the nonnative speaker with modified output: the listener may respond by speaking more slowly, enunciating more clearly, using simplified vocabulary, or paraphrasing (Gass & Varonis, 1984). Such modified interaction may be beneficial to the nonnative speaker and may help the nonnative speaker comprehend and communicate. On the other hand, a negative consequence may ensue. The accent may reduce the listener's ability to understand the nonnative speaker and may trigger negative, discriminating views of the speaker (Lippi-Green, 1997; Munro et al., 2003).

Research has shown that familiarity with a particular accent makes that type of accented speech easier to understand than speech with an unfamiliar accent (Gass & Varonis, 1984; Major, Fitzmaurice, Bunta, & Balasubramanian, 2002; Tauroza & Luk, 1997). This may be because "the phonetic characteristics of foreign-accented speech are highly systematic and quite consistent across talkers from the same native language background" (Bradlow & Bent, 2008, p. 708). Speech perception research has shown that listeners adapt to foreign-accented speech when repeatedly presented with it—meaning that repeated exposure to foreign-accented speech results in an increase in word- or sentence-level processing speed (Clark & Garrett, 2004) or increased accuracy in sentence recognition (Bradlow & Bent, 2008; Weil, 2001). However, the ease in comprehension due to familiarity may be mediated by the listeners' attitude toward the perceived accent—be it positive or negative. According to Major et al. (2002), it is often the case that language learners are stereotyped based on their accents, and they may be stereotyped as having

a lower social status. “Stereotypes regarding nonnative, accented speech seem to exist as perceptual constructs in the minds of both NSs and NNSs of English, and these attitudes have strong influences on listening comprehension” (Major et al., 2002, pp. 176-177). Thus, the relationship between accent and listening comprehension is complex. Furthermore, robust research on this relationship within the context of standardized assessment of oral speech samples by raters who are familiar with some, but not all accents in the speech sample pool is, as far as we know, extremely rare (see, however, Xi & Mollaun, 2009, reviewed above).

The studies of accent familiarity from the speech processing literature most relevant for this study are those focusing on judgments of nonnative speaker speech samples that differ in terms of accent. Scales, Wennerstrom, Richard, and Wu (2006) analyzed the accent perception of 37 English language learners and 10 American undergraduate students. All participants listened to an English passage read out loud by four people who were native speakers of, respectively, American English, British English, Mandarin Chinese, and Mexican Spanish. Results showed that the American students were best at identifying the American English and Mexican Spanish accents. Scales et al. speculated that the American students could easily identify the Spanish accents because they were learners of Spanish, and thus were familiar with the Spanish accent. Likewise, the English language learners whose native languages were Chinese and Spanish had higher accuracy rates in identifying the accents of speakers with whom they shared an L1. Scales et al. also found strong correlations between accent preference and ease of understanding. The participants liked an accent more if it was easy to understand. If they were unfamiliar with an accent, they were less able to identify or understand it, and tended to dislike it. These results are consistent with the results of matched-guise studies in which the same speaker is rated differently depending on the language he or she uses (Anisfeld & Lambert, 1964; Giles, 1971; Lambert, 1967; Lambert, Frankel, & Tucker, 1966; Lambert, Giles, & Picard, 1975; Lambert & Tucker, 1975). For example, Cargile and Giles (1998) found that native speakers of English evaluated Japanese-accented English speech samples more harshly than other, nonstandard accents of American English. They suggested that the evaluators’ perceptions of the Japanese social group as being negatively competitive may have influenced the evaluators’ perceptions. Thus, it appears that certain aspects of language may be thought of as being more “pleasant” than others, showing that raters’ attitudes may affect ratings.

These findings are different from those of others. As reported in Munro et al. (2006), Flege (1998) demonstrated that Chinese listeners scaled the accents of Chinese speakers in much the same way as native English speakers did. MacKay, Flege, and Imai (2006) found that Arabic listeners rated Italian-accented English in much the same way as native English listeners did. Munro et al. wrote, “these findings suggest that even listeners from very different L1 backgrounds might agree more than they disagree with respect to their perceptions of L2 speech” (p. 114). However, Munro et al. noted, these studies only had the listeners judge levels of accentedness of speech samples, and not their levels of intelligibility or comprehensibility. Thus, these studies’ results are inconclusive concerning whether listeners are biased when rating speech with accents.

Sometimes speech perceived as heavily accented is also perceived as completely comprehensible and intelligible (Brodkey, 1972; Derwing & Munro, 1997; Munro & Derwing, 1995; Smith & Bisazza, 1982; Smith & Rafiqzad, 1979), thus, “listeners often assign good comprehensibility ratings to speech samples that they have also rated as heavily accented” (Derwing & Munro, 2005, p. 386). Results only show that listeners from different L1 backgrounds can equally recognize that accents are there. Munro et al. attempted to address this issue by investigating the intelligibility of L2 speech that 40 listeners who were native speakers of Cantonese, Japanese, Mandarin, and English rated. The listeners listened to short English speech samples (4.5 to 10.5 seconds) from 48 speakers who were native speakers of Cantonese, Japanese, Polish, and Spanish. The listeners transcribed the speakers’ utterances (as a measure of comprehensibility) and rated the speech samples along accentedness and intelligibility scales. Results showed that the listener groups assigned accentedness and intelligibility ratings that correlated moderately to strongly (from .652 to .893). Correlations were lower in terms of comprehensibility (.470 to .855). Analyses that probed the effects of the listeners’ L1 on intelligibility, comprehensibility, and accentedness found that the listener groups differed in their comprehension and intelligibility ratings of the four different accents. But the results were not clear. For example, Japanese listeners found Japanese speakers more intelligible than any other listener group did. The Japanese listeners rated the Japanese speakers as easier to understand than the Cantonese speakers, but not the other groups. Munro et al. suggested that the correlations demonstrated strong similarities among listeners in how they respond to L2 speech. Munro et al. wrote,

This study has shown that there is a likelihood of a shared response to L2 speech, even among listeners from linguistically diverse backgrounds. Therefore, it offers no reason to doubt that oral test scores can have predictive value. Nevertheless, these findings need to be replicated in other work involving different listeners, different listening conditions, and different modes of evaluation. (p. 128)

One might question how “shared” the responses to the L2 speech actually were when one sees that the correlations ranged from .470 to .893 among the listener groups on the scales of intelligibility, comprehensibility, and accentedness. The generalizability of the results from the Munro et al. (2006) study to oral language testing may also be questioned because of the limited exposure the listeners had to the speakers’ speech (4.5 to 10.5 seconds) and because they made comprehensibility and accentedness judgments without reference to a carefully designed scoring rubric. Raters were also untrained. In standardized, oral proficiency testing, the average time listening to a speaker may be 20 minutes; speakers will respond to various tasks (which will result in diverse, oral language output); scoring rubrics are detailed (research has even shown that raters prefer more detailed rating scales and their use results in higher inter- and intra-rater reliability—Knoch, 2009); and raters must be thoroughly trained in how to apply the scoring rubric. Thus, not only does this work need replicating, but researchers need to carry out the replication within a setting more representative of true oral language testing contexts.

Xi and Mollaun (2009) conducted a recent study of rater behavior in an oral language testing context that focused on the effects of accent familiarity on the evaluation of students’ oral English performance. Twenty-six bilingual speakers of one or more Indian languages and English from India were trained to score the TOEFL iBT[®] Speaking test. Half participated in regular rater training, while the other half participated in rater training that included information on how to score native Indian-language speaker’s English speech samples. The goal was to see whether rigorous, specialized training could mitigate any bias raters may have because they are extremely familiar with the test takers’ L2 accent. The underlying assumption was that such raters might be disproportionately lenient or harsh when rating Indian-language-accented English. Using correlation and interrater reliability statistics, results showed that the bilingual raters were as reliable in their rating of Indian-language-accented speech as official ETS raters were. Specialized training did not increase the interrater reliability of the bilingual raters as a whole. It did appear, however, to make the bilingual raters more internally consistent and reliable in their

scoring of Indian-language-accented English, especially when the raters were less skilled overall. Qualitatively, the bilingual raters reported that their familiarity with the accent did not make it more difficult to fairly score Indian-language-accented English. They indicated that they were confident in scoring the English of speakers regardless of the speakers' L1. Contrarily, those who received the specialized training reported that they found it very useful, and that it made them more confident in scoring Indian-language-accented English. Thus, Xi and Mollaun recommended that Indian-language/English bilinguals who will rate Indian-language-accented English participate in specialized training. They noted anecdotally that such training may also help other non-Indian-language-speaking raters who have perceived difficulties in scoring Indian-language-speakers' English speech.

The Need for Further Research Into Rater Effects in Speaking Assessment

The studies reviewed thus far provide evidence that raters' different backgrounds, including their language backgrounds, can influence their rating behavior, and that rigorous training may help them be more internally consistent and reliable when they rate speech. But further research is needed for three reasons. First, prior studies in this area have mainly investigated the relationship between the raters' L1 background and the L1 of the test takers, not the raters' L2 background (which may match the L1 of the test takers). We do not know whether it is important to identify raters' L2s, nor do we know the extent to which raters' knowledge of certain L2s may affect the rating procedure, if at all. Secondly, the one study that has looked at this in the oral context (Xi & Mollaun, 2009) was rather small scale (26 bilingual raters were included in the study) and investigated raters whose knowledge of the test takers' L1 was native-like. A larger-scale study is therefore needed with the type of raters who may be more commonly found in large-scale rating programs: raters that have knowledge of the test takers' L1, but are not native-like speakers or users of the test takers' L1. Third, no previous studies have attempted to use both robust quantitative and qualitative methods to investigate the relationship among raters' L2 knowledge, the test takers' L1, and the raters' assignment of ratings. The mixed methods approach of the present study seeks to enhance understanding of raters' behaviors by investigating not only the ratings that raters assign in relation to the raters' L2 and the speakers' L1, but also how the raters view the individual speakers' accents and oral language performance.

Using Introspection to Study Rater Effects in Performance Assessment

Research in the area of rater characteristics and second or foreign language testing has often relied on quantitative methods. Studies that do not use qualitative measures often lack an important component, and that is the ability to tap mental processes that may be suggestive of some particular characteristics not easily detected in a purely quantitative study. Thus, there is a call for more analyses of qualitative data to investigate the way raters approach the task of rating (Johnson & Lim, 2009). Some studies have used qualitative verbal reports or mixed methods to study rater cognition, but these have mostly been in the context of essay rating (Cumming, Kantor, & Powers, 2002; Knoch, 2009; Lumley, 2006; Vaughan, 1991). For example, Knoch (2009) used a multifaceted Rasch analysis and semistructured interviews to investigate raters' use and perceptions of two different rating scales for writing. Fewer researchers have used introspective measures to investigate the process of rating oral proficiency tests. In the context of face-to-face oral proficiency interviews, Brown (2000, 2003) used retrospective verbal reports to investigate raters' reactions to oral test takers' performances and to determine why they awarded the ratings they did. She found that in oral interviews, examiners' personal questioning styles and feedback techniques impacted the quality of elicited speech and affected the way the examiners themselves viewed the test takers' communicative abilities. Pollitt and Murray (1996) used a type of verbal report to investigate what raters thought about their ratings of sets of oral performances. They found that raters differentially assessed speech according to the level of proficiency of the test taker: test takers who were highly proficient were judged in terms of content, whereas those who were less proficient were rated more on their accurate use of grammar. Kim (2009) used qualitative measures to investigate native and nonnative speakers' rating process. Raters who were native speakers and nonnative speakers of English were prompted to write their justifications for assigning ratings to oral English speech samples; the researcher used an inductive approach to analyze the justifications. Kim found that the nonnative speakers provided fewer comments and focused less on content accuracy. This, Kim reported, may have been due to the different cultures of the rater group—all nonnative speakers were in Korea, all native speakers were in Canada. None of these studies, however, investigated raters' reactions to accents or accent familiarity.

Verbal reporting is a special type of introspection, and stimulated recall, in which a participant, after performing a task, provides a report of his or her thought processes during that

task, is one type of verbal reporting that is noninvasive. The main advantage of the use of verbal report is that one can often gain access to processes that are unavailable by other means. Unlike written introspection (as used in Kim, 2009), verbal reports are often spontaneous stream-of-consciousness reports with little filtering. As such, stimulated recall is extremely effective in helping us uncover cognitive processes that are not evident through simple observation (Gass & Mackey, 2000). Information gleaned from stimulated recall can, in turn, support the reliability and validity of inferences we make from scores on an assessment instrument (Cohen, 1998). Ross (1997) commented that introspection can help test developers evaluate how their rating scales are working by investigating the extent to which the descriptions listed in the scale correspond to the ones that the test taker deployed. In other words, Ross argued that the rating scale must not only match test outcomes, but also the performance of test takers during the test-taking process, and he believed retrospection could help verify this. We would like to extend this argument. We believe that introspection can help researchers evaluate the effectiveness of their rating scales. Introspective data can be used to investigate the extent to which the descriptions listed on the rating scale correspond to the ones that the raters use during the rating process.

Research Question

In sum, a vast literature suggests that there are important variables that may impact the way raters evaluate speech. As Munro (2008) explained, speech perception and evaluation relate both to the stimulus properties (the SP component), that is, the linguistic properties of the speech independent of any affective interpretation, and listener factors (the LF component), which is the “human” or affective component, including the listener’s previous experiences with the language and his or her familiarity with it. But we do not know enough about the relative contributions of the SP and LF components to the listeners’ judgments of speech (Munro et al., 2006). As test designers embark on any test creation and implementation (and particularly those that have significant import and consequences for the test taker), it is imperative to identify those variables and determine how they might influence a rater’s decision-making process. Therefore, the following question guided this study: Are there certain groups of trained raters (grouped by their L2) who exercise differential severity, depending on the L1 of the test taker?

Methodology

Participants

One hundred and seven raters participated in this study. Rater background characteristics are presented in Table 1. Their ages ranged from 18 to 61, with a mean age of 22. Of these, 30 were male and 77 were female. They had a range of experience with an L2, in some cases being nearly bilingual and in some cases having familiarity of the L2 by virtue of studying and interacting with speakers of that L2. Eleven of the raters had studied Korean as an L2, 48 had studied Spanish, and 41 had studied Chinese. Seven other raters had studied German ($n = 3$), French ($n = 2$), Arabic ($n = 1$) or Japanese ($n = 1$). Of the 100 raters who had studied Spanish, Korean, or Chinese, 28 had studied the language for less than 2 years, 49 had studied for more than 2 years, and 23 were heritage speakers of the language, meaning that, to a varying extent, their immediate family members speak (or spoke) the language natively, they were (and/or are) exposed to the language in a family setting, and they had ethnic ties to native speakers of the language. Fifteen of the 107 raters had had significant ESL and/or EFL teaching experience. In all instances, this experience exceeded 1 year. Almost all ($n = 92$) raters were undergraduates at Michigan State University (MSU) with no prior rating experience. Eleven were graduate students at the same university. Four were recent graduates or affiliates of the university.

Materials

Background questionnaire. There were a number of materials that are relevant to this study. Of primary importance was the background questionnaire (see Appendix A). We closely followed Dörnyei's (2003) and Dörnyei and Taguchi's (2009) suggestions in constructing our background questionnaire. Through this questionnaire, we explored the language background of the participants, including languages spoken in the home, languages studied in school (when, where, how much), and language-use experience (living abroad, significant friendships). In addition, the questionnaire asked about teaching experience (degrees, experience abroad, ESL) and disciplinary background (degrees, including the field; undergraduate major; and minor; courses taken as graduate or undergraduate students). Part of the questionnaire included a self-assessment of the raters' proficiencies in languages with which they noted that they had experience.¹

Table 1***Rater Background Characteristics***

L2	N	Male	Female	Mean age	ESL/EFL teaching experience	Level of L2 experience		
						< 2 yrs	> 2 yrs	Heritage
1. Spanish	48	9	39	23.17	8	13	34	1
2. Korean	11	3	7	22.70	2	5	1	5
3. Chinese	41	18	23	19.87	2	10	14	17
4. Other	7	0	7	23.14	3	-	-	-
Total	107	30	77	21.91	15	28	49	23

Note. EFL = English as a foreign language, ESL = English as a second language, L2 = second language.

Test taker sound files. Because ETS has an extremely large database from which we could draw speech samples for this study, we requested and received from ETS a sample of sound files that were balanced in terms of test taker L1 (Spanish, Korean, Chinese) and average speaking test score (1-4). The sound files were from 72 individuals who took the TOEFL iBT in the fall of 2006. Of the 72 test takers, 24 had Spanish as their L1, 24 had Korean, and 24 had Chinese. Within each L1 group, there were 12 males and 12 females. Also within each L1 group, six test takers (three males and three females) were at each of the four levels of proficiency as specified by their overall TOEFL iBT Speaking test scores, based on the ratings that official ETS raters had previously assigned to these test takers' speech samples.² On the TOEFL iBT, the test takers recorded responses to six tasks, which we labeled A through F. Tasks A and B represent independent speaking tasks, Tasks C and D represent listening/reading/speaking tasks, and Tasks E and F represent listening/speaking tasks. Therefore, there were 432 ratable speech samples total. These test taker categories are listed in Table 2.

Training materials. We used the 4-hour, online training program for new ETS raters to which ETS provided us access. Raters also used copies of the official ETS scoring rubrics (Appendix B), which ETS gave us to use during the training session. We should note that the 107 raters in this study did not undergo rater certification, as official ETS raters must. Therefore, this study's population of raters differs considerably from official ETS raters in at least three ways: (a) their training was not as rigorous, (b) they were younger, and (c) they had not yet completed a 4-year college or university degree program. These differences in training and other characteristics will be discussed further below.

Table 2***Distribution of the 432 Sound Files by L1, Gender, and TOEFL iBT Test Score***

Language	Gender	
	Male	Female
Spanish		
Score of 4	3	3
Score of 3	3	3
Score of 2	3	3
Score of 1	3	3
Korean		
Score of 4	3	3
Score of 3	3	3
Score of 2	3	3
Score of 1	3	3
Chinese		
Score of 4	3	3
Score of 3	3	3
Score of 2	3	3
Score of 1	3	3
Total	36	36

Note. L1 = first language.

Data Collection Design

We employed two different but complementary methodological approaches in carrying out this study: we analyzed rating data using a multifaceted Rasch measurement (MFRM) approach, and we collected data on raters' thought processes using introspection (stimulated recall) and analyzed those data using an inductive approach. In what follows, we outline our data collection design for the MFRM analyses. A discussion of the introspective measures can be found in the Procedure section below.

We used the FACETS computer program (version 3.66.1; Linacre, 2009) to analyze the ratings that raters assigned to the TOEFL iBT speech samples. The multifaceted Rasch measurement (MFRM) model is an extension of the one-parameter Rasch model (Bond & Fox, 2001) and allows for the inclusion of many aspects, or facets, of the rating procedure (Bachman, 2004, chapter 2; Myford & Wolfe, 2003, 2004; Wigglesworth, 1994); in our case, we included seven facets, which were the raters, the raters' L2, the raters' level of knowledge of their L2, the raters' ESL/EFL teaching experience, the test takers, the test takers' L1, and the six speaking tasks (A through F). (In this study, we focus on the raters' L2, the raters' level of knowledge of

their L2, and the test takers' L1.) One issue that made it somewhat difficult to specify the design to use before actual data collection was that we did not know exactly what the background characteristics of our raters were going to be, although we recruited to obtain equal proportions of raters with L2 backgrounds in Spanish, Korean, and Chinese.

We carried out four MFRM analyses using several different combinations of variables, or facets, in MFRM models in order to learn as much as we could from our analyses about how the raters and test takers performed. We also conducted a bias interaction analysis using the FACETS computer program. We then used the results from that analysis to inform our qualitative analyses of data obtained from our stimulated recall interviews to try to understand the nature and sources of possible bias in raters' ratings.

Procedure

Raters completed 4 hours of rater training online on Day One in a computer lab. The raters then returned to the computer lab to complete 4 hours of online rating. The rating of speech samples for this study had to be completed within 3 days after the initial training. Each day that a rater rated began with four calibration exercises to reorient the rater to the scoring rubric. Each rater was paid \$90 for participating in the study.

We gave raters the option of being videotaped for 20 minutes while rating. If they agreed, they were invited to come the following day to participate in a 20- to 30-minute stimulated recall session prompted by the videotape of themselves rating. Twenty-six raters (11 male, 15 female, average age 21; two of the females had prior ESL teaching experience; seven were heritage learners) participated in the stimulated recalls and were paid an additional \$25 for their participation. (Table 3 presents the stimulated recall participants by their L2 background and gender.)

Table 3

L2 Background and Gender of the Stimulated Recall Participants

L2	Male	Female
Spanish	4	8
Korean	1	--
Chinese	6	5
Other	--	2
Total	11	15

Note. L2 = second language.

Due to logistical constraints, it was not possible to identify a common set of speech samples that all 26 raters would rate to use as the basis for the stimulated recall sessions. Rather, in our study, each rater rated 82 samples in approximately 4 hours, and we arbitrarily spent 20 to 30 minutes of that time videotaping for the stimulated recall session. Therefore, the 26 raters did not all comment on the same set of speech samples, nor did they necessarily comment on the same number of speech samples. Nonetheless, the stimulated recall sessions gave us an opportunity to explore the raters' thought processes and strategies in assigning scores with a strong and recently recorded stimulus (Gass & Mackey, 2007). The amount of structure involved was minimal. We did not lead or focus the raters as they carried out the stimulated recall task. We believe this helped render the recalls less susceptible to researcher interference.

During a stimulated recall session, the researcher had the rater watch the video of himself or herself rating. The raters were also provided with the scoring rubric they used while rating. The researcher paused the video from time to time and asked standard stimulated recall questions, such as, "What were you thinking about when rating just then?" or, "What were you thinking about at that time?" or, "What were you thinking when you were listening to the speech sample?" (If the rater did not know, he or she was encouraged to say so.) Raters were allowed and encouraged to stop the video whenever they remembered something in particular that they were thinking at the time of rating. Each stimulated recall session lasted approximately 30 minutes.

The study's data collection design, including the optional path raters could take of being videotaped while rating and participating in a stimulated recall session, is presented in Figure 1.

For the rating procedure, we used an incomplete block design with six forms. There were 432 speech samples total to be rated. Therefore, each of the six forms had a base of 72 tasks ($432/6 = 72$). We distributed the speech samples among the six forms so that there was an equal balance in regards to task type, the test takers' L1, holistic score level (based on prior ETS ratings of each speech sample), and gender.

After distributing the speech samples across the six forms, we chose 12 speech samples from across the forms to serve as anchor tasks (also referred to as "linking tasks"). The FACETS computer program would use the anchor tasks to link the raters. We selected these 12 anchor tasks by counter-balancing the speech samples according to form (two from each form), L1 background (four from each L1), holistic score level (three from each score category), and

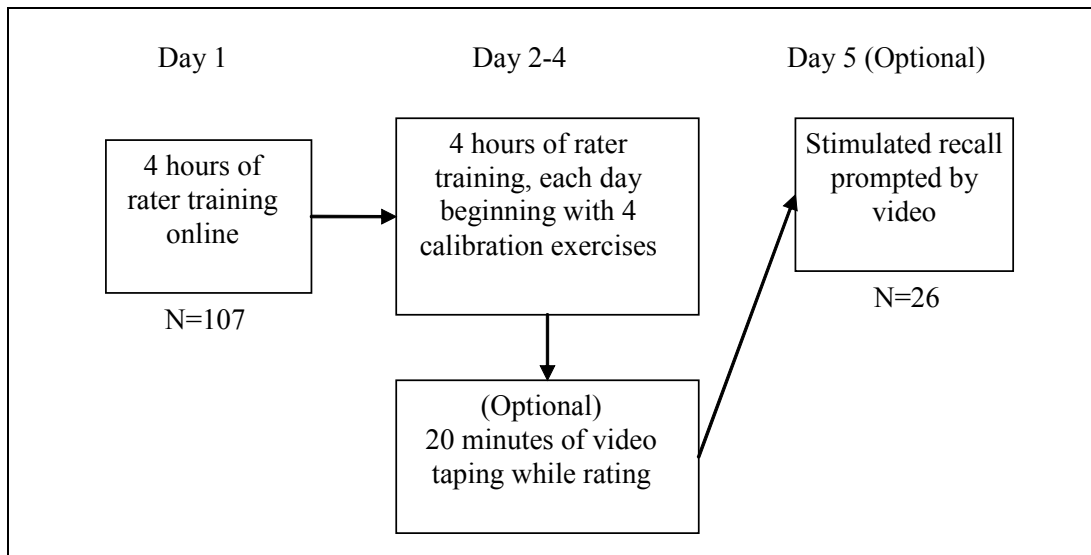


Figure 1. The study's data collection design.

gender (six from female test takers, six from male test takers). Every rater in the study rated these 12 anchor tasks, regardless of the form we assigned the rater to rate. Because we selected two anchor tasks from each of the six forms, any given form, therefore, had an additional 10 anchor tasks to be rated. That is, we took two anchor tasks from each form ($2 \times 6 = 12$), so that any one form had 72 speech samples plus 10 anchor tasks from the other forms ($72 + 10 = 82$). Thus, regardless of the rating form the rater received, each rater rated 82 speech samples. Our goal was to have at least 16 raters rate each form ($16 \times 6 = 96$ raters total) to provide sufficient data in order to calculate precise estimates of rater severity and test taker proficiency. Regardless of the form we gave the rater, he or she rated the speech samples in six blocks ordered sequentially by task: first the Task A speech samples, then the Task B speech samples, and so on. Within each block of tasks, the computer program randomized the presentation of the speech samples.

Data Analysis

Quantitative. The basic question for this study was whether there were any problematic interactions between the raters' L2 and the test takers' L1, which may have introduced unwanted construct-irrelevant variation into the rating procedure and, thus, possibly distorted the meaning of the resulting test scores and the inferences that test users might make from those scores. We used a MFRM approach to investigate the effects of seven variables, or facets, of the assessment

setting. Our test takers facet included 72 different test takers (the elements of the test takers facet). The raters facet was composed of 107 raters. The speaking tasks facet comprised six elements—the individual speaking tasks, A through F. Other facets included the test takers’ first language, L1 (Spanish, Korean, or Chinese); the raters’ second language, (L2 Spanish, Korean, or Chinese); the raters’ level of knowledge of their L2 (heritage speaker, more than 2 years’ experience, or less than 2 years’ experience); and the ESL/EFL teaching experience of the rater (more than 1 year vs. no experience or less than 1 year).

Using a MFRM approach, we analyzed the ratings the raters assigned to the speech samples in terms of group-level main effects for the facets included in the analyses. We used a rating scale model (RSM) because each task was rated on the same rating scale. (If individual items are rated on unique rating scales, a partial credit model, PCM, is used instead—see Linacre, 2000a.) We were able to separate out each facet’s contribution to the assessment setting and examine it independently of other facets to determine to what extent each facet was functioning as intended. When a MFRM analysis is run, the FACETS computer program analyzes the various facets simultaneously but statistically independently and calibrates them onto a single linear scale (i.e., the logit scale). The joint calibration of facets makes it possible to measure rater severity on the same scale as test taker proficiency and task difficulty. All facets of the rating operation are expressed in a common equal-interval metric (i.e., log-odds units, or logits). One can view the logistic transformation of ratios of successive category probabilities (log-odds) as the dependent variable, with various facets, such as test takers, tasks, and raters, conceptualized as independent variables that influence these log-odds. If the rating data show sufficient fit to the model, then researchers can draw useful, diagnostically informative comparisons among the various facets (as well as among the elements within a facet).

In this study, the multifaceted Rasch measurement model takes the following basic form:

$$\ln[P_{nij k} / P_{nij k-1}] = B_n - D_i - C_j - F_k \quad , \quad (1)$$

where

$P_{nij k}$ = probability of test taker n receiving a rating of k on task i from rater j ,

$P_{nij k-1}$ = probability of test taker n receiving a rating of $k - 1$ on task i from rater j ,

B_n = oral language proficiency of test taker n ,

D_i = difficulty of task i ,

C_j = severity of rater j , and

F_k = difficulty of receiving a rating of k relative to a rating of $k - 1$.

For this study, we carried out four separate MFRM analyses using the FACETS computer program. First, to determine whether test taker first language (L1) subgroups differed in their average levels of oral language proficiency, we used the following MFRM model:

$$\ln[P_{nijkm} / P_{nijkm-1}] = B_n - D_i - C_j - T_m - F_k , \quad (2)$$

where

T_m = the test taker's first language (L1) m .

Second, to determine whether raters with more second language experience tended to rate any more severely or leniently on average than raters with less second language experience, we used this MFRM model:

$$\ln[P_{nijko} / P_{nijko-1}] = B_n - D_i - C_j - K_o - F_k , \quad (3)$$

where

K_o = the rater's level of knowledge of their second language o .

Third, to determine whether raters having more than 1 year of ESL/EFL teaching experience tended to rate any more severely or leniently on average than raters with no experience (or less than 1 year of ESL/EFL teaching experience), we used this MFRM model:

$$\ln[P_{nijkp} / P_{nijkp-1}] = B_n - D_i - C_j - E_p - F_k , \quad (4)$$

where

E_p = the rater's ESL/EFL teaching experience p .

Fourth, we conducted a bias interaction analysis to determine whether rater L2 subgroups exercised differential severity when rating various test taker L1 subgroups of speech samples. In order to carry out this analysis, we added several facet terms and an interaction term to Equation 1. Shown below is the multifaceted Rasch measurement model we used to investigate the rater L2 x test taker L1 interaction:

$$\ln[P_{nijklm} / P_{nijklm-1}] = B_n - D_i - C_j - T_m - R_l - I_{lm} - F_k, \quad (5)$$

where

P_{nijklm} = probability of test taker n in first language subgroup m receiving a rating of k on task i from rater j in second language subgroup l ,

$P_{nijklm-1}$ = probability of test taker n in first language subgroup m receiving a rating of $k-1$ on task i from rater j in second language subgroup l ,

B_n = oral language proficiency of test taker n ,

D_i = difficulty of task i ,

C_j = severity of rater j ,

R_l = rater second language (L2) subgroup l ,

T_m = test taker first language (L1) subgroup m ,

I_{lm} = interaction between rater second language subgroup l and test taker first language subgroup m , and

F_k = difficulty of receiving a rating of k relative to a rating of $k-1$.

I_{lm} is a summary statistic that indicates the degree to which the ratings of rater second language subgroup l for test taker subgroup m differ from the expected ratings of rater subgroup l for test taker subgroup m . (The expected ratings are derived from a MFRM model that includes facets for raters and test takers but no rater subgroup x test taker subgroup interaction term.) FACETS computes the bias interaction term using a two-stage calibration process. In the first stage, the computer program estimates all parameters except I_{lm} . In the second stage, FACETS anchors all parameters except I_{lm} to the values estimated during the first stage and then obtains parameter estimates and standard errors for I_{lm} .

Qualitative. We were interested in analyzing the actual ratings the raters assigned, but we were also interested in the reasons for assigning particular ratings. Two of the researchers (Winke and Gass) used stimulated recall data to support the quantitative results and to provide further insight into what raters were thinking about as they were rating. It is one thing to see quantitatively that one group rates familiar speech differently than they rate other speech; it is another to be able to delve into the reasons for this. The latter information can come from stimulated recall data. At the outset of the study, we anticipated that there might be comments

relating to other variables that we had not considered in this study; thus, during the stimulated recall we were open to further questions or comments if the raters wished to elaborate.

To analyze the data from the stimulated recalls, we followed the guidelines set out in Mackey and Gass (2005). We followed an inductive approach, in which themes and patterns emerged from the data. We were aided in our thematic analysis for coding and interpreting the stimulated recall data by the use of the qualitative analysis software package QSR NVivo 8. After we transcribed all stimulated recall audio files, we entered the data into QSR NVivo 8. We read the data segments (a segment is a participant's single response during the stimulated recall—there were 260 in our data) and subsequently independently grouped them into various themes and patterns. We then discussed the themes and patterns we had identified. Our discussion led to splitting some themes and consolidating others. We discussed and agreed upon specific names for, and operationalizations of, the themes. We then reread and recoded the data segments using the consolidated themes. Agreement was 91% (236 out of 260). Finally, we discussed the 24 data segments that we coded differently until we could reach a consensus on each segment's classification.

Before discussing results related to our research question, we first present results from our analyses regarding the fit of the data to the Rasch model and the assumption that the data is unidimensional. Use of measures derived from a multifaceted Rasch measurement analysis requires that the data demonstrate sufficient fit to the model and that the test measure a single unidimensional construct. That is, the test scores must reflect the measurement of a single, unitary ability or trait (in this case, speaking ability), and each item must contribute meaningfully to the measurement of this one latent trait (Henning, 1992). In this study, we examined the fit statistics for the tasks in our data set to determine whether the rating data exhibited sufficient fit to the model (Bonk & Ockey, 2003). This represents a “fit-only” approach to unidimensionality testing, one of three main approaches for assessing unidimensionality discussed in the literature (Tennant & Pallant, 2008). In addition, we examined the point-measure correlation coefficients to investigate data fit, as will be explained below after we discuss the fit statistics for the tasks.

Wright and Linacre (1994) suggested that a reasonable range of mean-square infit and outfit values for judge-rated items (or tasks) when agreement is encouraged is 0.4 to 1.2. As can be seen in Table 4, all the mean-square infit values for the tasks fall within this range (i.e., their infit mean-square values range from .88 to 1.2), and only one of the mean-square outfit values is

out of the range, but only slightly (i.e., Task B, with a mean-square outfit value of 1.25). Additionally, estimated task discrimination indices in the range of 0.5 to 1.5 indicate reasonable fit to the Rasch model (Linacre, 2000b). As reported in Table 4, all the estimated task discrimination indices are within this range, suggesting good model-data fit. The final column in Table 4 reports the point-measure correlation for each task. This is a measure of the degree to which the ratings that raters assigned test takers on a particular task were correlated with the test takers' proficiency measures. The point-measure correlations are all within the range of .82 to .88, indicating that higher ratings on each task correspond to higher overall scores.

Table 4

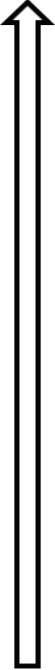
Item Facet Summary Statistics (Sorted by Infit Mean-Square Values)

Task	Obsv. raw score average	Fair-M average	Difficulty measure (in logits)	Model SE	Infit mean square	Outfit mean square	Est. discr.	r_{pm}^a
B	2.7	2.75	-0.38	0.05	1.20	1.25	0.72	0.82
F	2.6	2.68	-0.17	0.05	1.00	1.06	1.00	0.86
A	2.7	2.71	-0.25	0.05	0.99	1.02	1.01	0.86
E	2.7	2.76	-0.41	0.05	0.95	0.93	1.05	0.86
C	2.5	2.39	0.62	0.05	0.93	0.89	1.13	0.88
D	2.4	2.4	0.58	0.05	0.88	0.95	1.13	0.88

^a Point-measure correlation.

It should be noted, however, that model-data fit criteria can also be applied to the raters and their ratings. Normally, in a test validation study, raters (or even items) that do not fit the model are removed from the dataset, and then the model-data fit is evaluated again without them. Some of the infit mean-square values for raters are provided in Table 5, and it is evident that some of the raters had values above the recommended 1.2 maximum (for example, rater 136 had an infit mean-square index of 1.42). But the purpose of this study was not to provide validation evidence to support inferences made from scores on this assessment. Rather, we were explicitly interested in the “noise” in the data stemming from the raters, especially systematic variance that may be explained through bias interaction analysis. Thus, we deemed our rater data, which includes, by common categorization, misfitting data, worth investigating. In other words, our plan was to conduct a bias interaction analysis to identify particularly problematic interactions and then to use reasoning to try to gain an understanding of the nature of those problematic interactions.

Table 5***Differences in Severity/Leniency That Selected MSU Raters Exhibited***

	Rater ID	Number of ratings	Raw score average	Fair average	Rater severity measure (in logits)	SE	Infit mean-square index
Most severe	117	76	2.1	2.06	1.52	.22	.91
	79	76	2.4	2.21	1.12	.22	1.17
	136	76	2.3	2.23	1.04	.21	1.42
	93	76	2.3	2.25	.99	.22	.87
	121	76	2.4	2.30	.86	.21	.76
	90	76	2.3	2.31	.84	.21	1.01
	123	76	2.3	2.31	.82	.22	.81

	150	76	2.6	2.61	.03	.21	.72
	138	76	2.6	2.61	.01	.21	1.33
	68	76	2.6	2.62	.00	.22	1.29
	111	76	2.7	2.62	.00	.21	.89
	101	76	2.6	2.62	-.01	.22	.62

	112	76	2.8	2.89	-.77	.22	.85
	130	76	2.9	2.91	-.83	.22	.84
	129	76	2.8	2.95	-.97	.22	1.13
	125	76	3.0	3.08	-1.34	.23	1.03
	106	76	3.0	3.09	-1.38	.22	1.13
	Most lenient	86	76	3.0	3.11	-1.44	.23
174		76	3.0	3.19	-1.66	.22	.99
	Mean	76	2.6	2.61	.00	.22	.99
	SD	.00	.20	.20	.57	.00	.24

Note. Mean and standard deviation are for all 107 raters, not just those shown in the table.

MSU = Michigan State University.

Results

The results section is divided into two parts. The first part presents results from the multifaceted Rasch measurement analyses (quantitative). The second part presents results from the stimulated recall (qualitative) analyses.

Quantitative Results

Before presenting the results that answer the research question, we first provide a brief introduction to the process of interpreting the FACETS output. In particular, we focus on the variable map that is perhaps the single most important and informative piece of output from the computer program. It enables us to view all the facets of the analysis at the same time. The map (Figure 2) not only displays all facets of the analysis, but also summarizes key information about each facet. The map highlights results from more detailed sections of the FACETS output for test takers, raters, tasks, and the rating scale.

The FACETS computer program calibrates the test takers, the raters, the speaking tasks, and the rating scale so that all facets are positioned on the same scale, creating a single frame of reference for interpreting the results from the analysis. The scale is in log-odds units, or “logits,” which, under the model, constitute an equal-interval scale with respect to appropriately transformed probabilities of responding within particular rating scale categories.

1. The first column in the map displays the logit scale. Having a single frame of reference for all the facets of the rating process facilitates comparisons within and between the facets. The scale ranges from 6 logits to -6 logits.
2. The second column displays the scaled scores that ETS uses to report scores to test takers (Educational Testing Services, 2008). ETS averages the six ratings that two raters assign to each test taker (for a total of 12 ratings), and a single scaled score between 0 and 30 is reported. In our test taker sample, no test takers who received average ratings below 1 were included. Therefore, the MSU raters did not assign any zeros. The MSU raters were trained on and used a 4-point rating scale, from 1 to 4 (see Appendix B). Thus, for our data, possible scaled scores included 8, 9, 10, 11, 13, 14, 15, 17, 18, 19, 20, 22, 23, 24, 26, 27, 28, 29, and 30.³
3. The third column displays estimates of test taker proficiency on the speaking test. These are single-number summaries on the logit scale of each test taker’s tendency to receive low or high ratings across raters and speaking tasks. We refer to these as *test taker proficiency measures*. Higher scoring test takers appear at the top of the column, while lower scoring test takers appear at the bottom. Each star represents one

- test taker. In Figure 2, the highest scoring test taker had a proficiency measure of 5.76 logits. The lowest scoring test taker had a proficiency measure of -5.71 logits.
4. The fourth column displays the MSU raters in terms of the level of severity or leniency each exercised when rating the oral responses to the six tasks. Because more than one rater rated each test taker's responses, raters' tendencies to rate test takers' oral responses higher or lower on average could be estimated. We refer to these as *rater severity measures*. In this column, each star represents three raters. Each dot represents one or two raters. More severe raters appear higher in the column, while more lenient raters appear lower. In Figure 2, the harshest rater had a severity measure of 1.52 logits, while the most lenient rater had a severity measure of -1.66 logits.
 5. The fifth column compares the six speaking tasks in terms of their relative difficulties. Tasks appearing higher in the column were more difficult for the test takers. That is, it was more difficult for the test takers to receive high ratings on these tasks than on those tasks appearing lower in the column. Tasks C and D were the most difficult for the test takers, while the other tasks proved easier.
 6. The sixth column displays the 4-point rating scale that raters used to score test taker responses to each of the six tasks. The horizontal lines across the column indicate the point at which the likelihood of getting the next higher rating begins to exceed the likelihood of getting the next lower rating for a given task. For example, when we examine Figure 2, we see that test takers with proficiency measures from about -6 logits through -3 logits were more likely to receive a rating of 1 than any other rating. Test takers between about -3 and 0 logits were more likely to receive a rating of 2. Those between about 0 and 3 logits were more likely to receive a 3. And those between about 3 and 6 logits were more likely to receive a 4.

At the bottom of Figure 2 we provide the means and standard deviations of the distributions of estimates for test takers, raters, and tasks. When conducting a MFRM analysis involving these three facets, it is customary to center the rater and task facets, but not the test taker facet. By centering facets, we established the origin of scale.

(1) Logit Scale	(2) ETS Scaled Score	(3) Test Taker	(4) Rater	(5) Task	(6) Rating
		High Scores	Severe	Difficult	
+ 6		+ *	+	+	+ (4) +
		**			
+ 5		+	+	+	+
	28-30				
		**			
+ 4		+ *	+	+	+

+ 3	-----	+ *****	+	+	+ --- +
		**			
		**			
+ 2		+ *****	+	+	+
		*			
	22-27	*	.		3
		*			
+ 1		+ **	+ *	+	+
		**	****.		
		**	***.	C-integrated D-integrated	
		**	****.		
* 0	-----	* *****	* *****.	*	* --- *
		**	****.	A-independent F-integrated	
		**	*****	B-independent E-integrated	
		**	***		
+ -1		+ ***	+ .	+	+
		***	.		
	14-20	**	.		2
		*	.		
+ -2		+ *	+	+	+
		*			

+ -3	-----	+ *	+	+	+ --- +
		**			

+ -4		+ *	+	+	+
		**			
	8-13				
+ -5		+ *	+	+	+
		*			
		*			
+ -6		+	+	+	+ (1) +
		Low Scores	Lenient	Easy	
Mean		.28	.00	.00	
S.D.		2.88	.57	.43	

Figure 2. Variable map from the FACETS analysis of the data.

As shown in Figure 2, the distribution of rater severity measures was much narrower than the distribution of test taker proficiency measures. In Figure 2, the test taker proficiency measures showed a 12-logit spread, while rater severity measures only showed about a 3-logit spread. This is important because it suggests that the impact of individual differences in rater severity on test taker scores was likely to be relatively small. By contrast, if the range of rater severity were wider than the range of test taker proficiency measures, the impact of individual differences in rater severity on test taker scores would have been much greater. The particular raters who rated individual test takers would have mattered more, and there would have been a stronger need to adjust test taker scores for individual differences in rater severity in order to minimize such biasing effects.

FACETS also reports a test taker separation ratio (G), which is a ratio scale index comparing the “true” spread of test taker proficiency to its measurement error (Fisher, 1992). Adequate separation is important in cases in which a test produces scores that test users employ to separate test takers into categories defined by their performance (Stone & Wright, 1988). The MSU raters should be using the scoring rubric such that the test takers are divided into distinct strata of proficiency. Using the test taker separation ratio, we calculated the number of statistically distinct proficiency strata into which the test (and raters, in this case) succeeded in separating test takers by using the formula $(4G+1)/3$.⁴ The test taker separation index was 14.12, which indicated that this sample of test takers could be separated into 14 statistically distinct levels of proficiency. The reliability of the test taker separation index was 1.00, indicating that the test takers were very reliably separated in terms of their levels of proficiency (something we already knew based on the pre-rating scores that ETS provided to us). This finding suggests that the MSU raters used the scoring rubric adequately to distinguish between test takers of differing levels of proficiency.

We used FACETS to produce a measure of the degree of severity each rater exercised when rating test takers’ speech samples from the TOEFL iBT Speaking test. Table 5 shows a portion of the output from the FACETS analysis summarizing the information the computer program provided about the MSU raters. The raters are ordered in the table from most severe to most lenient. The higher the rater severity measure, the more severe the rater. To the right of each rater severity measure is the standard error estimate, indicating the precision with which the rater was measured. Other things being equal, the greater the number of ratings a severity

estimate is based on, the smaller the standard error. The rater severity measures for the raters ranged from 1.52 logits to -1.66 logits, a 3.18 logit spread.

A more substantive interpretation of rater severity is obtained by examining each rater's mean rating. Even though all MSU raters did not rate all test takers, each rater's mean rating should not have been influenced much by the sample of test taker performances that he or she rated (assuming that the ETS raters' ratings were accurate). Remember that ETS raters pre-rated each test taker's set of six speech samples, and we received each test taker's average score across the six tasks. We used those holistic ratings to distribute the speech samples across the six forms. Each MSU rater rated the speech samples from one of the six forms. We balanced the forms so that each form contained an equal number of speech samples to which the ETS raters assigned ratings of 1, 2, 3, or 4. We also balanced the forms in terms of gender and L1. Thus, the mean ratings of the raters should have been fairly equal, regardless of which form a particular rater rated. Therefore, we can compare the mean ratings of any two MSU raters, even though each rated a different (yet overlapping) set of test takers.

Another way of comparing the raters is to examine the mean rating for each rater once it has been corrected for the deviation of the test takers in each rater's sample from the overall test taker mean across all raters and tasks. This *fair average* allows one to determine the extent to which the mean ratings of raters differed after having taken into account the particular sample of test takers that each rater evaluated. Thus, the fair average adjusts a rater's raw mean rating based on whether or not he or she rated a batch of speech samples with a disproportionately high number of high or low scoring test takers.

For example, the mean rating of the most severe MSU rater was 2.1, and the rater's fair average was 2.06 (see Table 5). By comparison, the mean rating of the most lenient MSU rater was 3.0, and that rater's fair average was 3.19. This means that, on average, the two most extreme MSU raters assigned ratings that were 0.9 raw score points apart when we compare their mean ratings, and 1.13 raw score points apart when we compare their fair averages. We could report either as the spread of rater severity, but reporting the fair average spread is particularly warranted when interpreting our results because not all raters rated all tasks (Wolfe & Dobria, 2008). With a fair average spread > 1 (or one level of proficiency on a 4-point scale), we can see that the range of rater severity is fairly wide.

By looking at the Facets output, we can see who, exactly, is rating more severely than others and who is rating more leniently than others. As indicated by the rater severity measures reported in Table 5, we can see that the following were the extreme raters at each end: raters 117, 79, 136, 93, and 121 were rating more harshly than others; raters 174, 86, 106, and 125 were rating more leniently than others.

FACETS also provides a chi-square test of the hypothesis that all 107 MSU raters exercised the same degree of severity when rating the test takers' speech samples. The overall resulting chi-square value was statistically significant; the chi-square value of 724.0 with 106 degrees of freedom ($p = .00$) signified that at least two of the raters did not exercise the same level of severity when evaluating the test takers' speech samples. At the very least, the most severe rater and the most lenient rater were significantly different.

The rater separation index was 2.44. This suggests that there were about two-and-a-half distinct strata of rater severity in this sample of 107 raters. (This index is calculated using the formula $(4G + 1)/3$, where G is the rater separation ratio, and statistically distinct levels of rater severity are defined as severity strata that are three standard errors apart, centered on the mean of the rater sample. See Wright and Masters [2002] for more information.)

The reliability of the rater separation index was .86. This index provides information as to how well one can differentiate among the raters in terms of their levels of severity. It is the Rasch equivalent of a KR-20 or a Cronbach alpha test reliability statistic (i.e., the ratio of true variation to observed variation for the elements of a particular facet). It is not a measure of interrater reliability, which is a measure of how similar the raters are. Nor is it an indication of how well the assessment is functioning (Linacre, 2010). Rather, rater separation reliability is a measure of how reproducibly different the rater severity measures are (Linacre, 2010). The most desirable results would have been to have a rater separation reliability index close to zero, which would have suggested that raters were interchangeable, exercising very similar levels of severity. If the rater separation reliability were 1.0, then that would indicate that the raters were completely different in terms of their levels of severity and were not at all interchangeable. In the context of our study, the rater separation reliability of .86 suggests that there was evidence here of unwanted variation in rater severity that could have affected test taker scores. We will take this information into consideration when we discuss the results.

To answer the research question (Are there certain groups of trained raters [grouped by their L2] who exercise differential severity, depending on the L1 of the test taker?), we conducted a bias interaction analysis to determine whether raters were rating in a similar fashion the speech samples that native (L1) speakers of Chinese, Korean, and Spanish produced, or whether some subgroups of raters appeared to exhibit a bias toward (or against) speech samples that any of the test taker L1 subgroups produced. Specifically, we wanted to find out whether any of the rater subgroups (categorized by L2) showed evidence of exercising differential severity/leniency, rating any specific test taker L1 subgroup more severely or leniently than expected, or whether each rater subgroup’s average level of severity/leniency was invariant across test taker L1 subgroups.

Table 6 provides summary statistics related to overall test taker L1 subgroup differences in performance for the study’s 107 raters. From left to right, the columns in Table 6 present the test taker L1 subgroup, the sum of the MSU raters’ ratings of speech samples for that L1 subgroup, the count of the number of rated speech samples that contributed to the observed raw score, the observed raw score average (i.e., the observed raw score divided by its count), the test taker L1 subgroup’s average proficiency measure (in logits), and the standard error associated with that measure.

Table 6
Test Taker L1 Subgroup Measurement Report

Test taker L1 subgroup	Observed raw score	Observed count ^a	Observed raw score average	Average proficiency measure (in logits)	Model SE
Korean (<i>n</i> = 24)	6,421	2,532	2.5	-0.06	0.02
Chinese (<i>n</i> = 24)	6,539	2,543	2.6	-0.02	0.02
Spanish (<i>n</i> = 24)	6,761	2,525	2.7	0.08	0.02
Mean	6,573.7	2,533.3	2.6	0.00	0.02
SD	141	7.4	0.1	0.06	0.00

Note. Fixed (all-same) chi-square = 28.6; *df* = 2; *p* = .00. L1 = first language.

^a The count of the number of rated speech samples that contributed to the observed raw score.

Based on our prior knowledge that the three test taker L1 subgroups should not have differed in terms of their average proficiency, the results from the fixed chi-square test are useful in determining whether the 107 raters exhibited a group-level differential severity/leniency effect when rating speech samples from the three subgroups. The results are shown at the bottom of Table 6. The chi-square value of 28.6 with 2 degrees of freedom was statistically significant ($p = .00$), indicating that at least two of the average proficiency measures for the three test taker L1 subgroups were statistically significantly different. This was not expected. When choosing the speech samples to be included in this study, ETS purposely selected three sets of test takers who, according to the ratings the ETS raters assigned their speech samples, exhibited the same average proficiency. However, when the MSU raters rated these test takers' speech samples, the average proficiency measures for the three test taker subgroups were not equal. These results suggest that some MSU raters may have exercised differential severity/leniency when rating test takers in the different subgroups.

Based on the ETS raters' ratings of the test takers' speech samples, we hypothesized that all three subgroups of test takers should have approximately the same average proficiency measure after accounting for measurement error. In Table 7, which is based on the data presented in Table 6, we see that this is not the case. The average proficiency measure for L1 Korean test takers was $-.06$ logits ($SE = .02$). The average proficiency measure for L1 Chinese test takers was $-.02$ ($SE = .02$). The average proficiency measure for L1 Spanish test takers was $.08$ logits ($SE = .02$). The difference between the average proficiency measures was $.04$ logits for the Korean and Chinese test taker subgroups, $.14$ logits for the Korean and Spanish subgroups, and $.10$ logits for the Chinese and Spanish subgroups. It appears that the raters overall were more lenient toward test takers who had Spanish as an L1, and more severe toward test takers who had Korean or Chinese as an L1. However, it should be noted that other researchers posit that differences between subgroup performances of less than $.30$ logits are usually not substantively meaningful (Engelhard & Myford, 2003), thus suggesting that the differences found here between the Spanish (on the one hand) and Korean and Chinese (on the other hand) L1 subgroups are not generally indicative of a strong overall group-level rater bias. Nonetheless, some evidence of bias appears to be present, which warrants statistical comparisons.

Table 7***Differences in L1 Test Taker Subgroups' Average Proficiency Measures***

Test taker L1	Average proficiency measure (SEM)	Mean differences		
		Korean	Chinese	Spanish
Korean	-.06 (.02)	--	.04	.14*
Chinese	-.02 (.02)		--	.10*
Spanish	.08 (.02)			--

Note. L1 = first language.

* $p = .01$.

We performed three t tests to compare the average proficiency measures for the three test taker L1 subgroups.⁵ In two of the comparisons the differences were statistically significant at the .01 level. The Spanish L1 and Chinese L1 subgroup average proficiency measures were significantly different, $t(87) = 3.52, p = .00$, with L1 Spanish speakers receiving higher ratings on average than L1 Chinese speakers. The Spanish L1 and Korean L1 subgroup average proficiency measures were significantly different, $t(57) = 4.44, p = .00$, with L1 Spanish speakers receiving higher ratings on average than L1 Korean speakers. However, the Chinese L1 and Korean L1 subgroup average proficiency measures were not significantly different, $t(50) = 2.36, p = .02$.

When the raters are grouped by their L2s, the picture is more complex. Table 8 contains summary statistics for each of the three rater subgroups' (grouped by L2) ratings of each of the three subgroups of test takers (L1 Spanish, Korean, and Chinese). The first column (Observed raw score) displays the sum of the rater subgroup's ratings for that test taker subgroup. The second column (Expected raw score) displays the sum of that rater subgroup's expected ratings for that test taker subgroup based on the calibrations from the main FACETS analysis (i.e., an analysis that only looks at main effects of the variables included in the measurement model, not any interaction effects). The third column (Observed count) is the number of estimable responses involving this particular test taker subgroup and this particular rater subgroup (e.g., in the first line, we see that raters who had Korean as an L2 rated 284 speech samples that native speakers of Korean produced). The fourth column (Observed-expected average) is the observed raw score minus the expected raw score divided by the observed count. This is the size of the bias calculated in terms of raw score units. The fifth column (Bias size) displays the size of the bias in

logit units relative to the rater subgroup's overall severity measure (e.g., the overall level of severity that the raters with Korean as an L2 exercised when rating all test takers' speech samples, .02 logits, is shown in the far right column of Table 8). A *t* statistic (column seven) accompanies each bias size and is used with the degrees of freedom (the observed count, which is the number of cases involved minus 2) and the *p* value to determine whether the interaction between the two subgroups was statistically significant. The *p* values are listed in the final column. The ninth column reports *r* values (i.e., effect size estimates). Two of the interactions were statistically significant at the .01 level: (a) the raters with Spanish as an L2 were significantly more lenient toward test takers who had Spanish as an L1, and (b) the raters with Chinese as an L2 were significantly more lenient toward test takers who had Chinese as an L1. In other words, the raters with L2 backgrounds in Spanish or Chinese tended to be more lenient when rating test takers who had the L1 of the language they had studied. However, in both cases the effect sizes were small, each accounting for less than 1% of the variance in the ratings.

In Figure 3, we can see more clearly the differences in the levels of severity with which the rater L2 subgroups rated the test taker L1 subgroups. The figure displays bias interactions between test taker L1 and rater L2. The test taker L1 subgroup observed raw score averages for each rater L2 subgroup are plotted along the Y axis, and the rater L2 groups appear along the X axis. This figure reveals that overall, regardless of their L2, raters were more lenient when rating speech samples of test takers with Spanish as an L1 than when rating speech samples of test takers with Chinese or Korean as an L1.

However, when we examine the specific interactions between rater L2 and test taker L1, the findings are somewhat more nuanced: (a) the raters with Spanish as an L2 ($n = 48$) were significantly more lenient toward the Spanish-native-speaking test takers than the raters with either Korean ($n = 11$) or Chinese ($n = 41$) as an L2, and (b) the raters with Chinese as an L2 were significantly more lenient toward the Chinese-native-speaking test takers than the other raters. It should be noted here that while we found evidence of significant biases toward the native speakers of the raters' L2 for the raters with Spanish and Chinese as their L2, the rater group with Korean as an L2 was rather small ($n = 11$). Thus, absence of evidence of bias in the ratings that this rater group assigned may reflect the small rater sample size.

Table 8***Bias/Interaction Report for Test Taker L1 Subgroups and Rater L2 Subgroups***

Observed raw score	Expected raw score	Observed count ^a	Observed - expected average	Bias size ^b	Model SE	<i>t</i>	<i>df</i>	<i>r</i>	Test taker L1	Rater L2	Rasch logit measure ^c	<i>p</i>
718	689.4	284	0.1	0.09	0.06	1.61	283	.10	Korean	Korean	0.02	0.108
3,288	3,167.8	1,214	0.1	0.09	0.03	3.29	1,213	.09	Spanish	Spanish	-0.02	0.001 *
2,690	2,588.5	1039	0.1	0.09	0.03	2.99	1,038	.09	Chinese	Chinese	0	0.003 *
718	695.7	283	0.08	0.07	0.06	1.26	282	.07	Chinese	Korean	0.02	0.209
2,620	2,539.3	1,035	0.08	0.07	0.03	2.38	1,034	.07	Korean	Chinese	0	0.018
3,083	2,993.7	1,213	0.07	0.07	0.03	2.43	1,212	.07	Korean	Spanish	-0.02	0.015
3,131	3,055.7	1,221	0.06	0.06	0.03	2.05	1,220	.06	Chinese	Spanish	-0.02	0.041
2,768	2,706.8	1,042	0.06	0.05	0.03	1.81	1,041	.06	Spanish	Chinese	0	0.071
705	690.5	269	0.05	0.05	0.06	0.84	286	.05	Spanish	Korean	0.02	0.402
2,191.2	2,125.3	844.4	0.08	0.07	0.04	2.07			Mean	(Count: 9)		
1,064.4	1,032.5	406.6	0.02	0.02	0.01	0.75			SD			

Note. Fixed (all-same) chi-square = 43.7; *df* = 9; *p* = .00. L1 = first language. L2 = second language.

^a The count of the number of rated speech samples that contributed to the observed raw score. ^b The higher the bias size, the more biased the rater subgroup was toward the test taker subgroup (i.e., the more likely those raters were to give higher ratings than expected to speech samples from that test taker subgroup). ^c The rater subgroup's overall severity measure (i.e., the average level of severity the rater subgroup exercised when rating all test takers' speech samples).

**p* = .01.

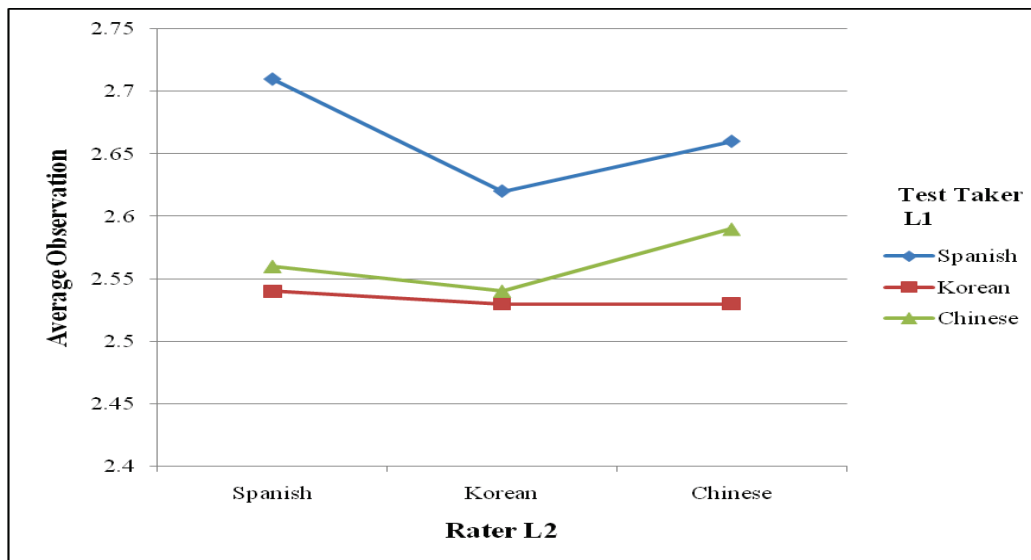


Figure 3. Bias/interaction analysis specified by test taker L1 and rater L2.

We also looked at whether raters with more L2 experience tended to rate any more severely or leniently on average than raters with less L2 experience. Results from the FACETS analysis are presented in Table 9. According to the results from the fixed-effect chi-square test, none of the differences among the average severity measures for the three levels of L2 experience were statistically significant, $\chi^2(2, N = 3) = 4.1, p = .12$. (See Table 9.) Raters with significant L2 exposure (i.e., heritage speakers and those with more than 2 years experience) did not tend to rate any more or less harshly on average than raters with little L2 exposure (i.e., less than 2 years experience).

We also looked at whether raters having more than 1 year of ESL/EFL teaching experience tended to rate any more severely or leniently on average than raters with no experience (or less than 1 year of ESL/EFL teaching experience). The results are presented in Table 10. The results from the fixed-effect chi-square test indicated that the average levels of severity that the two subgroups of raters exercised were not significantly different, $\chi^2(1, N = 2) = 0.20, p = .63$. However, it should be noted that the two rater subgroups are unequal in sample size, with only 15 of the 107 raters having more than 1 year of ESL/EFL teaching experience. Thus, results from this analysis are tentative in suggesting that raters with more than 1 year of ESL/EFL teaching experience do not tend to rate any more or less harshly on average than raters with less than 1 year of ESL/EFL teaching experience, or no experience.

Table 9***Rater L2 Subgroup Measurement Report***

Rater L2 experience subgroup	Observed raw score	Observed count ^a	Observed raw score average	Average severity measure (in logits)	Model <i>SE</i>
Little exposure (i.e., less than 2 years; <i>n</i> = 28)	5,479	2,128	2.6	0.02	0.02
Significant exposure (i.e., more than 2 years; <i>n</i> = 49)	9,628	3,724	2.6	0.01	0.02
Heritage speakers (<i>n</i> = 23)	4,614	1,748	2.6	-0.04	0.02
Mean	6,573.7	2,533.3	2.6	0	0.02
<i>SD</i>	2,188.4	856.1	0	0.03	0

Note. Fixed (all-same) chi-square = 4.1; *df* = 2; *p* = .12. L2 = second language.

^a The count of the number of rated speech samples that contributed to the observed raw score.

Table 10***Rater ESL/EFL Teaching Experience Subgroup Measurement Report***

Rater ESL/EFL teaching experience subgroup	Observed raw score	Observed count ^a	Observed raw score average	Average severity measure (in logits)	Model <i>SE</i>
No experience, or less than 1 year (<i>n</i> = 92)	17,143	6,612	2.6	0.01	0.01
More than 1 year (<i>n</i> = 15)	2,578	988	2.6	-0.01	0.03
Mean	9,860.5	3,800	2.6	0	0.02
<i>SD</i>	7,282.5	2,812	0	0.01	0.01

Note. Fixed (all-same) chi-square = .2; *df* = 1; *p* = .63. EFL = English as a foreign language, ESL = English as a second language.

^a The count of the number of rated speech samples that contributed to the observed raw score.

Qualitative Findings

The 26 raters who participated in the stimulated recall sessions produced 20,997 words total. The average number of words per rater was 808—the minimum words spoken was 214 (rater 113); the maximum was 1,816 (rater 142). Table 11 presents the coding categories that emerged from the data. The “Count” column in Table 11 shows the number of raters who made a comment related to the specific coding category, the total number of references that the raters made regarding the coding category, and the number of words that the raters used in discussing the particular category. For example, 15 raters commented on the accents of the test takers. These 15 raters made 29 individual comments on accent, which totaled 1,754 words. If appropriate, we further coded a comment on a test taker’s accent as positive or negative. We coded six raters’ comments on the test takers’ accents as positive (e.g., they commented that the accent was good, contributed to a higher rating, or gave a good impression of the test taker) and 12 as negative (e.g., the raters commented that the accent made scoring or listening difficult). We judged two of the references to accent as neither positive nor negative—this is reflected in the references column where the total number of comments on accent is 29, yet only 27 were coded as positive or negative.

The three topics that the raters discussed most frequently (i.e., what the raters claimed they were thinking about while rating) were the test taker’s ethnic accent, the task at hand, and aspects pertaining to the test taker’s voice (i.e., voice volume and the quality of the voice). At least 15 of the 26 raters discussed each of these topics. Raters also reported that they thought about the L1 of the test takers, responded emotionally to the test takers, and experienced difficulties in assigning ratings. They also noted technical problems that sometimes made rating difficult. In addition, four of the seven raters who were heritage language learners mentioned that certain speech samples triggered thoughts about their own heritage status while they were rating them.

Through our analysis of the qualitative data, we have demonstrated that raters face a number of challenges when rating speech samples. Some of these challenges are independent of the raters’ own personal backgrounds. For example, the qualitative results revealed that 9 of the 26 raters who participated in the stimulated recalls found that some of the test takers had technical difficulties in recording or using their microphones, which made rating problematic. However, some of the challenges were related to the individual raters’ background characteristics,

Table 11***Summary of Coded Data From Stimulated Recalls***

Coding category	Count		
	Raters	References	Words
1. Test taker's accent	15	29	1,754
1.1. Positive	6	9	296
1.2. Negative	12	18	1,280
2. Task	15	19	1,260
2.1. Task A; describe a book	13	17	1,119
2.2. Task B; television, good or bad?	1	1	91
2.3. Task F; discuss science experiment	1	1	50
3. Test taker's voice	15	28	1,557
3.1. Volume of voice	13	18	992
3.1.1. Quiet	12	15	873
3.1.2. Loud	2	2	100
3.2. Quality of voice	8	9	551
3.2.1. Positive	1	1	27
3.2.2. Negative	8	9	555
4. Test taker's L1	12	19	1,246
5. Affect (how rater felt while listening)	10	19	1,328
5.1. Sorry	8	13	981
5.2. Mad, upset, or angry	2	2	129
5.3. Happy	2	2	101
6. Scoring difficulty	9	11	778
7. Technical problem	9	11	556
8. Rater's heritage status	4	7	840

including the L2 experience of the raters. Our research question asked whether there were groups of trained raters (grouped by their L2) who exercised differential severity depending on the L1 of the test taker. From the results of our bias interaction analyses, we found that in certain cases, when there was a match between the raters' L2 and the test takers' L1, some raters tended to rate more leniently. Here, we turn to the qualitative data to see if there is any confirmation of this result. We believe there is. We identified three themes that appear to relate to, and partially

confirm, the relationship between rater L2 and test taker L1. These themes were borne out when raters discussed the following: (a) the test taker's accent, (b) the test taker's L1, and (c) their own heritage status as part of the rating process. The following sections present findings related to each of these three themes.

The test taker's accent. Raters often commented that they thought about the test takers' accents while rating. They reported that accents affected their listening processes and may have potentially influenced the ratings they assigned. For example, Rater 111 (female, learner of Spanish) noted that she could identify Asian accents even though she had not studied an Asian language herself. She stated the following when asked what she was thinking while rating a native speaker of Chinese.

Example 1. I was having a hard time understanding this speaker, and um, I think going through my head in this particular one was I was trying to not be biased on accents because, um, just for me as the listener, I was thinking um, some of the more like Asian accents are harder for myself to understand...

Likewise, Rater 131 (female, Chinese heritage learner) commented on her awareness of her potential bias related to the accent of a test taker. As one more familiar with Asian accents, she stated that she was afraid she was too lenient toward those with Asian accents.

Example 2. I was trying to guess where the speaker was from, and um, like it, I thought maybe that was maybe playing into some sort of prejudice that I had but I wasn't sure. But I was worried that since I had worked more with Asian students that I would give them higher scores because I was more comfortable with their accents.

Rater 60 (male, Chinese learner), Rater 94 (male, Chinese learner) and Rater 142 (male, heritage learner of Chinese) stated similar concerns to those of Rater 131. They all stated that they felt they would be (or were) too lenient toward those who had accents (L1s) with which they were familiar. Their comments are presented, respectively, below.

Example 3. I was always, because I remember reading try and be as objective as possible, don't be sympathetic because you know how hard ch, uh, English is to learn, and I was just wondering if that was influencing my scoring. Because yeah, I was born in the Philippines, and my parents aren't native speakers, so I'm used to working through accents. My parents' accents are thick at times, so I wondering I might be too generous,

and sometimes I was wondering am I overcompensating by being too harsh, so I never really knew if I was being perfectly objective.

Example 4. I'm used to listening to more of the eastern accent. I wondered how that tied into my theory of of uh that sheet [scoring rubric]. You know, I'm a Chinese for my major, so I know a Chinese accent better so when they're speaking, I'm like oh, I can hear clearly.

Example 5. Um, I knew, I knew right off the bat that I was probably going to be slanted when it comes to how to grade because I think I told you before, I'm first generation. So I've grown up hearing this kind of English, and I know that it's been my job for the past 18 years to fill in, to fill it in to make it sound more English-sounding so my mind already knows how to do that. So every time I heard it I already had, I already had to make sure I was going to catch myself if I started to hear their voice, put it in mine, and then send it back out saying this is what she really meant, I know what she really meant, she just can't say it right now because maybe she was never taught that particular word. So, in my mind the whole time there was just a fear that I was going to slowly, slowly begin to be very, very easy on each individual when they spoke. Especially I was definitely very scared that if I ever heard an Asian sounding accent that I was going to be very, very, very easy on them, on their voice.

Other raters also discussed their preferences for (or against) strong accents or particular accents, and this, they admitted, affected their rating. For example, Rater 128 (female, Chinese heritage learner, listening to a native speaker of Spanish) commented that she paid a lot of attention to accent when rating, and Rater 129 (another female Chinese heritage learner, listening to a native speaker of Korean) noted that she most likely rated a particular Korean test taker lower based on her accent. Their respective comments are presented below.

Example 6. Um, I think I remember she was, she was stumbling a lot over her pronunciation. Um, but I just kinda felt bad for her but I don't know, I guess there was a degree of people. Some that talked that sounded like native speakers actually and some that like, you know they're from a different country because of their heavy accent. You know, you can tell kinda what's what. Cuz like, she knew what she wanted to say but there's still the second area, the accent. I just thought it was terrible.

Example 7. Her accent was really confusing me actually, and I had to actually pay like a little bit more attention than I needed to it. So like, right there I was like, um I don't think she's gonna get a 4, she's probably gonna end up getting a 3.

The test taker's L1. Twelve of the 26 raters provided detailed comments concerning how they often thought about the L1s of the test takers. They guessed what the test takers' L1s might be. They wondered from which countries the test takers originated. They thought about these characteristics of the test takers at various points throughout the task of listening to a recorded, oral speech sample. For several of the raters, it seemed they pondered about the L1 of the test taker whenever they listened to a new speech sample. The explicit pinpointing of a test taker's L1 might imply that these raters were sensitive to qualifications within the speech sample that they could identify. And for some raters, it seemed perplexing to not be able to figure out the test taker's L1. For example, their comments included, "I think I might've been wondering where he was from," and "Um, I remember thinking, I did not understand where she's from, well, when she started out," and "Well, I figured she was from a Spanish-speaking country!" Some questioned their desire or need to identify the L1 of the test taker and wondered whether their ability or non-ability to do so affected their rating processes. For example, one rater mentioned, "I couldn't tell what his L1 was but then I thought, well, that's probably a good thing." A similar comment was made by Rater 131 (female, Chinese heritage learner), who stated the following:

Example 8. I noticed not just with him but with a lot of them, almost everything that I rated, um, I had to keep stopping myself from trying to guess which country they were from. Because I kept trying to guess which country they were from because I'm an ESL teacher. That was going on in the back of my mind while I was thinking.

The rater's heritage status. Out of the 26 raters who participated in stimulated recall sessions, seven were heritage learners. Four of those seven mentioned that when listening to speech samples, characteristics within the speech samples made them think about their own experiences as heritage language learners. For example, Rater 142 (male, Chinese heritage learner) noted the following:

Example 9. [Listening to a Chinese native speaker] And I caught myself because I started really sympathizing with this man because he sounded just like how my father speaks.

And I was like, he speaks a lot like how my father does where sometimes he puts tonals in where you're not supposed to put a tone. But I was like, well, that doesn't stop me from understanding what you're trying to say. It might be a little annoying after awhile, but it doesn't mean that I didn't know what you were saying to me the whole time.

Likewise, Rater 65, a heritage learner of Spanish, noted paying particular attention to test takers whose L1 matched the language she was learning as a heritage language learner. She said the following:

Example 10. [Listening to a Spanish native speaker] I also wondered if that's what I sound like, when I am speaking, because I can perceive her she speaks Spanish, or Italian, as a first language. So I start listening to her, if the language was typical. I am self-conscious about the way I sound, so I was very interested in listening to her a little bit more to see about her, to see about where her, where she makes mistakes, not grammatical mistakes but pronunciation mistakes.

To summarize, a number of the native-English-speaking raters at MSU reacted to the L1s of the test takers. The results from our qualitative analysis indicate that some raters' prior L2 learning experiences may have influenced their ratings, perhaps in certain cases interacting with the L1s of the test takers. A number of the raters commented that they thought about the test takers' L1 accents, and that they frequently tried to identify the test takers' L1s. They acknowledged that they were most successful in doing so when their own L2 was the L1 of the test takers. In many cases, it appeared that the test taker's accent seemed to affect the way these raters perceived the quality of the test taker's speech. Some might view this as "off rubric" thinking (i.e., when rating the speech samples, taking into consideration a criterion that the scoring rubric designers may have regarded as inappropriate to include as part of the rubric). Lastly, some of the raters who were heritage language learners appeared to sense the impact of the test takers' L1 on their own rating behavior. Their interactions with the test takers' L1 at times brought out strong emotive aspects and made them think about their own language learning processes and linguistic backgrounds. These interactions seemed to be a natural part of the raters' listening processes, regardless of whether they affected the assignment of ratings.

The results from the analyses of the quantitative data answered the research question affirmatively. Certain groups of trained raters (grouped by their L2) exercised differential

severity, depending on the L1 of the test taker. The results from the analysis of the qualitative data helped to posit some possible explanations for those severity differences: Some raters tried to discern the L1 of the test takers, which may be a natural response to hearing non-native-speaker speech, especially for those accustomed to hearing it. Accent familiarity, in turn, may have affected comprehension. And some raters may perceive an emotional connection to test takers based on the test takers' foreign accent or perceived L1, which may result in the assignment of a biased rating.

Discussion

Many participants are involved in the evaluation of speech samples, including rubric designers, rater trainers, and raters who use scoring rubrics to assign ratings. All of these individuals may leave their mark on the test process, inadvertently affecting test scores in some way. In spite of this, we expect the scores that test takers receive from standardized tests of oral proficiency to be accurate and appropriate for the test users' purposes (Luoma, 2004). In fact, test performance is normally attributed to the performance of the test taker alone (McNamara, 2001). But we cannot ignore the fact that the interpretation of the scoring rubrics and rating criteria may “act as *de facto* [emphasis added] test constructs” (McNamara, Hill, & May, 2002, p. 229). Thus, it is critical to understand how raters interpret and apply scoring rubrics, and whether raters use criteria not on the rubrics when assigning ratings.

Our study was unique because we investigated raters' prior L2 learning experiences to gain an understanding of how raters' and test takers' language backgrounds may influence the ratings that raters assign. In the past, researchers carrying out studies of raters assessing oral proficiency have thus far only investigated how shared L1s among raters and test takers may influence raters' ratings (Brown, 1995; Fayer & Krasinski, 1987; Kim, 2009; Xi & Mollaun, 2009)—raters in those studies were bilingual, that is, highly proficient, near-native or native speakers of the language being tested. We expanded the research paradigm to determine whether raters who are non-native-like learners of the test takers' L1 would assign ratings that were significantly higher or lower than expected. Effectively, we asked whether a rater's language learning background is a legitimate factor for investigation in rater bias studies, even when the raters did not acquire the language to an advanced or native-like level. Results from this study seem to suggest that a rater's language learning background is indeed a very legitimate factor for investigation. The results from our bias interaction analysis indicate that the MSU raters were

more lenient in assigning ratings to test takers whose L1 they (the raters) had studied. Based on this study's outcome, we suggest that, when shared with the test taker's L1, a rater's L2 may exert just as much an influence on the rating process as a bilingual rater's L1. Thus, testing programs may need to be aware of their raters' L1 (bilingual) and L2 backgrounds. Moreover, the implications for rater training programs that Xi and Mollaun (2009) suggested may apply even when the raters involved have only studied (and not mastered) the test takers' L1.

Our main recommendation from this study is that testing programs for speaking assessment may want to consider including in their rater training programs specific modules that cover various sources of bias related to rater and test taker backgrounds that might impact raters' ratings of speech samples. The goal of such training would be to attempt to sensitize rater trainees to these potential sources of construct-irrelevant variance, such as the test takers' accents, and by extension, the raters' familiarity with the test takers' non-native-like encoding and word- and sentence-level processing (Bradlow & Bent, 2008; Clark & Garrett, 2004; Weil, 2001). The training could also help rater trainees understand how those various sources of bias might inadvertently (or perhaps intentionally) influence the ratings they assign. As the speech processing literature suggests and this study's stimulated recall data evidence, when raters can identify the test takers' ethnic accent, they also tend to be able to more easily understand the test takers' speech. This may be why some of the raters who participated in the stimulated recall sessions in this study mentioned that they often, if not always, tried to identify the accents of the test takers. Being able to do so most likely correlated with better comprehension of the speech stream. This aligns with results from Derwing and Munro (2005), who found that even heavily accented speech can be very comprehensible. Moreover, we believe, based on our data, that this may be the case when the accented speech is familiar at some level.

Lumley (2002) argued that rating is certainly possible without training, but training and recalibration sessions are essential so raters can adequately develop a sense of how the institution or test administrators interpret the scoring rubric. During training sessions, the raters are informed about how the test designers envision the scoring rubric should be used, and they often practice using the rubric to assign ratings to various previously rated benchmark language samples. According to Elder, Barkhuizen, Knoch, and von Randow (2007),

Training has been found to attenuate extreme differences between raters in terms of severity, to increase the self-consistency of individual raters by reducing random error

and also to counteract individual biases in relation to the various dimensions of the rating situation (i.e., task, scale and candidate) (McIntyre, 1993; Weigle, 1994a; 1994b; 1998; Wigglesworth, 1993). (p. 38)

Lumley noted that through training, raters learn how to justify their rating decisions. They learn how to describe, in terms that the testing organization or institution establishes and are found on the rubric, why they think a language sample is at a certain level on the scale. Lumley (2002, 2006) further noted that even with training, there may still be a tension between how the rater perceives the language sample and his or her efforts to apply the scoring rubric. This may have been evidenced by Xi and Mollaun (2009) who found that raters felt better about rating after receiving specialized rater training to deal with biases—biases that actually did not show up in the ratings in their study.

The results from these rater training studies would seem to suggest that the raters in this study may not have received enough training. First, some of the MSU raters in this study evidenced biases, which suggests that they may be good candidates for more training. It could be that the 4-hour rater training program was not enough to instill in these raters a sense of how we wanted them to interpret the scoring rubric. With more training, we could have done a better job of informing the raters about how they should employ the scoring rubric and carry out the rating process. Perhaps they did not know how to justify their rating decisions. It could be that better trained raters would not volunteer information that they listened for accents, wondered about the L1s of test takers, or thought about their own heritage language experience while listening. In this sense, perhaps it was beneficial to conduct this study with relatively untrained raters to uncover natural biases that training may need to address.

We have further evidence that the raters in this study might have benefitted from additional training. The FACETS analysis uncovered unwanted variation in rater severity: the rater separation index was 2.44, indicating that there were about two-and-a-half distinct strata of rater severity in our sample of 107 raters. Thus, the raters in this study did not all exercise the same degree of severity when rating the test takers' speech samples. More ideal would have been to have raters whose severity measures clustered more tightly around 0 logits on the FACETS map (Figure 2). A very narrow distribution of rater severity measures would indicate that the raters were assigning ratings in a similar fashion, which is what any testing program would want. If the raters in this study were exercising different levels of severity, they might not have been

focusing enough on the criteria specified in the scoring rubric and might have been considering other criteria that the scoring rubric did not include. Thus, the MSU raters in this study were not functioning interchangeably—some of the variation in the levels of severity they exercised may have resulted from the relatively short rater training program we provided, and the fact that the raters themselves were not experienced or “certified” to rate in any way. They did not have to pass a qualifying test to be allowed to rate in this study. Although raters must normally possess college degrees, most of the raters in this study were undergraduates. However, our reasons for selecting raters from a college undergraduate pool were twofold. First, we needed to recruit raters who we knew would fulfill our criteria of having experience learning the L1s of the test takers; and second, we wanted to control the languages with which those raters would have had experience (Chinese, Korean, and Spanish). Thus, we employed a convenience sample, which worked well to investigate the question at hand, but the nature of the sample also limits the generalizability of the study’s findings. First and foremost, the results of this study cannot be directly generalized to trained and certified ETS raters. It is very possible that with more extensive rater training, the process of certification, and experience (educational, teaching, and professional rating experience) raters may be less likely to exhibit biases.

Nonetheless, we believe that the results from this study are important because they reveal that raters’ prior L2 study may influence rating behavior, just as raters’ additional L1s may influence rating behavior. Additionally, the study shows that raters’ L2s and test takers’ L1s may interact, resulting in raters exercising differential levels of severity. Whether these language-background-related biases would persist with extensive rater training needs investigation. In any case, when training raters to rate nonnative speech samples, test developers would be wise to consider the many factors that determine how speech is evaluated, especially in a high-stakes testing situation (Chalhoub-Deville & Turner, 2000). This study reveals that raters’ prior L2 study may be one of those factors.

Conclusion

Because rating necessarily involves human judgment, validation of scores requires gathering various lines of evidence that present a sound and convincing validity argument. In helping to build that argument, we sought to determine whether there were factors that might influence raters’ ratings of speech samples that were not part of the scoring rubric raters were to apply. The results from our bias interaction analysis revealed that rater and test taker background

characteristics may exert an influence on some raters' ratings. We found that when there is a match between the test taker's L1 and the rater's L2, some raters may be more lenient toward the test taker and award the test taker a higher rating than expected.

After identifying this unexpected bias, we sought some possible explanations for why it occurred. We wanted to know whether the raters themselves would comment on this bias, and whether they were aware that such a bias existed. The qualitative results from the stimulated recall data made us aware that some of the raters did indeed feel that such a bias could occur and perhaps did exist—some mentioned their problems and difficulties in rating, and these often centered around their problems dealing with the accents of the test takers. Some of the raters even wondered out loud if they were more lenient toward the test takers who shared the same ethnicity as they did, and other raters commented on difficulties rating test takers whose accents were unknown, particularly “thick,” or difficult to decipher. However, these two criteria—test taker accent, and test taker L1—are not criteria that appear in the scoring rubric. Rater training typically does not cover these criteria or provide guidelines concerning what raters should do when they come across a test taker whose accent is particularly “foreign” or, on the other end, extremely familiar due to the rater's experience in learning the native language of the test taker (cf. Xi & Mollaun, 2009).

In this sense, accent and the identification of ethnicity through accent may be elephants in the language testing rating room—they are issues that may be omnipresent in oral assessment, but are mostly unaddressed. The results from our bias interaction analysis indicate that biases toward test taker L1 exist; the results from our stimulated recall analysis indicate that some raters may be aware of their own personal biases along these lines and may be uneasy about them; however, the scope and magnitude of the problem are not known because raters are not normally sorted and categorized by their prior L2 experience, and bias interaction studies of this type are rare.

There are limitations to this study that we must address. First, it is possible to question the extent to which the data from the stimulated recall sessions are valid and reliable. For example, are the reports consistent with the actual behavior of raters when rating? Is investigating rater cognition, that is, the decision-making behaviors that raters use when they evaluate language samples (Cumming, et al., 2002), worthwhile? We assume so, mainly because various researchers have shown that verbal reports such as think-alouds and stimulated recall are

reliable measures, and that results obtained using verbal reports do correspond with actual behavior (see Brown, 2005; Gass & Mackey, 2000; Kim, 2009; Lumley, 2006). Many researchers have found that introspective analysis of performance can yield important information on the cognitive processes that underlie performance on certain tasks. Even so, the validity of the raters' claims should be interpreted with caution. In this study we reported on stimulated recall data, but we were concerned that some of the raters' comments gathered through stimulated recall appeared at times to go beyond their thought processes at the time of rating (see Examples 3, 5, and 6).

Additionally, we assume that the 26 stimulated recall participants' data is representative of the entire set of 107 raters, but this is purely speculative. Our speculation is as follows: we found systematic effects with respect to score assignments based on rater L2 background (the quantitative findings), and we found issues with score interpretation along the lines of rater L2 background knowledge (the qualitative findings), and thus we posit that the former may stem from the latter. However, the comments themselves do not provide information as to how well the individual raters were actually rating. Future qualitative inquiry of this nature could be more fine-grained and include information on the actual scores the raters assigned to the particular speech samples being mentioned. Multifaceted Rasch analysis could be used to identify biased raters, and then those raters could be targeted for an interview or stimulated recall session to uncover the nature, intentionality, and/or awareness level of their biases. Nonetheless, the stimulated recall data in this study should be seen as intriguing triangulation, and we believe they provide a valuable supplement to the rating data.

Second, the raters in this study were young as well as inexperienced in the rating process. Several of the raters were 18; the average age was 22. However, the study's weaknesses were also its strengths—being able to cull the participants from a large undergraduate pool allowed us to tightly control the language background variables in the rater population and match that to the test takers' L1s, which makes this study particularly valuable. Past studies investigating the possible influences of raters' prior language experience on the rating process have employed small rater sample sizes (Johnson & Lim, 2009; Kim, 2009). This study, in contrast to Johnson and Lim in particular, has shown that with a larger sample of raters, discernable patterns of language-background related biases can appear.

Next Steps

Using qualitative and quantitative methods, we identified possible sources of construct-irrelevant variance in the ratings of oral speech samples. When raters know, to varying degrees, the L1 of the test takers and can discern the L1 of the test takers through the test takers' ethnic accents, that knowledge may influence their ratings. Replication studies are needed to determine how pervasive this problem is; and if this challenge persists, specific guidelines for dealing with the bias need to be provided in rater training modules and recalibration sessions. In particular, the prior language study of raters needs to be documented and considered when monitoring rater performance so that this type of rater bias can be more easily identified and addressed.

Another finding of this study is that overall, raters were more lenient toward test takers who had Spanish as their native language and more harsh toward test takers who had Korean or Chinese as their native language. In our study we did not uncover concrete reasons to explain this finding, but we suspect it may be unique to this set of raters (who were all native speakers of English, which is more closely related to Spanish) and our particular study's context. American students in Michigan most likely have more opportunities to hear Spanish-accented English than Korean- or Chinese-accented speech; thus, overall, they may be more familiar with it—which follows our theory that they then may rate it more leniently. However, our musings need empirical backing.

We collected additional data during this study that we did not analyze. The computer program recorded the raters' reaction time and time spent on listening and rating. For example, it recorded the time of the rater's first click to listen to an audio sample. Also, it recorded the amount of time spent listening to the speech sample (raters were able to rewind or fast-forward an audio file using the audio player's control features) and the number of times a rater played each audio file. When the rater entered the speech sample's score, the computer recorded the time. These data are available for subsequent post-hoc analyses and could reveal interesting findings in relation to rater hesitation, amount of time spent on rating, and the overall consistency and severity/leniency of raters.

References

- Andrich, D. (1988). *Rasch models for measurement: Quantitative applications in the social sciences*. London, England: Sage.
- American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anisfeld, E., & Lambert, W. (1964). Evaluational reactions of bilingual and monolingual children to spoken language. *Journal of Abnormal and Social Psychology*, *69*, 89-97.
- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.
- Barkaoui, K. (2007). Participants, texts, and processes in ESL/EFL essay tests: A narrative review of the literature. *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, *64*(1), 99-134. doi: 10.3138/cmlr.64.1.099
- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.
- Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89-110.
- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707-729.
- Brodkey, D. (1972). Dictation as a measure of mutual intelligibility: A pilot study. *Language Learning*, *22*(2), 203-220.
- Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1-15.
- Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. In R. Tulloh (Ed.), *IELTS research reports* (vol. 3, pp. 49-84). Canberra, Australia: IELTS Australia.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, *20*(1), 1-25.
- Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt, Germany: Peter Lang.

- Cargile, A. C., & Giles, H. (1998). Language attitudes toward varieties of English: An American-Japanese context. *Journal of Applied Communication Research*, 26(3), 338-356.
- Chalhoub-Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge Certificate Exams, IELTS, and TOEFL. *System*, 28(4), 523-539.
- Clark, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America*, 116(6), 3647-3658.
- Cohen, A. D. (1998). Strategies and processes in test taking and SLA. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 90-111). Cambridge, England: Cambridge University Press.
- Connor-Linton, J. (1995). Crosscultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes*, 14, 99-115.
- Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7(1), 31-51.
- Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, 86(1), 67-96.
- Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19(1), 1-16.
- Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, 39(3), 379-398.
- Dewey, D. P. (2004). A comparison of reading development by learners of Japanese in intensive domestic immersion and study abroad contexts. *Studies in Second Language Acquisition*, 26(2), 303-327.
- Dörnyei, Z. (2003). *Questionnaires in second language research: Construction, administration, and processing*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language Research: Construction, administration, and processing* (2nd ed.). New York, NY: Routledge.
- Educational Testing Services. (2008). *Converting rubric scores to scaled scores*. Retrieved January 17, 2008, from http://www.etsliteracy.com/Media/Tests/TOEFL/pdf/Converting_Rubric.pdf

- Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 24(1), 37-64.
- Engelhard, G., Jr., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition program with a many-faceted Rasch model*. New York, NY: College Board.
- Esling, J. H., & Wong, R. F. (1983). Voice quality settings and the teaching of pronunciation. *TESOL Quarterly*, 17(1), 89-95.
- Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, 37(3), 313-326.
- Fisher, W. P., Jr. (1992). Reliability statistics. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 6(3), 238.
- Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America*, 76, 692-707.
- Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gass, S. M., & Mackey, A. (2007). *Data elicitation for second and foreign language research*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gass, S. M., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, 34(1), 65-89.
- Giles, H. (1971, October 14). Our reactions to accents. *New Society*, 713-715.
- Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H. W. Dechert & M. Raupach (Eds.), *Interlingual processes* (pp. 229-244). Tuebingen, Germany: Gunter Narr Verlag.
- Hamp-Lyons, L., & Zhang, B. W. (2001). World Englishes: Issues in and from academic writing assessment. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 101-116). Cambridge, England: Cambridge University Press.
- Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, 9(1), 1-11.
- Hill, K. (1997). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 275-290). Jyvaskyla, Finland: University of Jyvaskyla.

- International Language Testing Association. (2005). *Code of ethics for ILTA*. Retrieved from <http://www.iltaonline.com/code.pdf>
- Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, 26(4), 485-505.
- Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, 26(2), 187-217.
- Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, 26(2), 275-304.
- Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly*, 26(1), 81-112.
- Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19(1), 3-31.
- Lambert, W. (1967). A social psychology of bilingualism. *Journal of Social Issues*, 23(2), 91-109.
- Lambert, W., Frankel, H., & Tucker, G. (1966). Judging personality through speech: A French-Canadian example. *International Journal of the Sociology of Language*, 158(4), 305-321.
- Lambert, W., Giles, H., & Picard, O. (1975). Language attitudes in a French-American community. *Linguistics and Education*, 158(4), 127-152.
- Lambert, W., & Tucker, G. R. (1975). White and Negro listeners' reactions to various American-English dialects. In J. Dillard (Ed.), *Perspectives on Black English* (Vol. 4, pp. 369-377). The Hague, The Netherlands: Mouton.
- Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.
- Linacre, J. M. (2000a). Comparing "partial credit" and "rating scale" models. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 14(3), 768.
- Linacre, J. M. (2000b). Item discrimination and infit mean-squares. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 14(2), 743.
- Linacre, J. M. (2009). *FACETS Rasch-model computer program* (Version 3.66.0) [Computer software]. Chicago, IL: Winsteps.com.
- Linacre, J. M. (2010). *A user's guide to FACETS Rasch-model computer programs. Program manual 3.67.0*. [Software manual]. Available at www.winsteps.com

- Lippi-Green, R. (1997). *English with an accent: Language ideology and discrimination in the United States*. New York, NY: Routledge.
- Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19(3), 246-276.
- Lumley, T. (2006). *Assessing second language writing: The rater's perspective*. Frankfurt, Germany: Peter Lang.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12(2), 238-257.
- Luoma, S. (2004). *Assessing speaking*. Cambridge, England: Cambridge University Press.
- MacKay, I. R. A., Flege, J. E., & Imai, S. (2006). Evaluating the effects of chronological age and sentence duration on degree of perceived foreign accent. *Applied Psycholinguistics*, 27(2), 157-183.
- Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Major, R. C., Fitzmaurice, S. M., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly*, 36(2), 173-190.
- Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing*, 22(1), 59-92.
- McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, 18(4), 333-349.
- McNamara, T., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics*, 22, 221-242.
- Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. H. Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 193-218). Philadelphia, PA: John Benjamins.
- Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning*, 48(1), 73-97.

- Munro, M. J., Derwing, T. M., & Burgess, C. S. (2003). The detection of foreign accent in backwards speech. In M. J. Solé, D. Resasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 353-538). Barcelona, Spain: Universitat Autònoma de Barcelona.
- Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, 28(1), 111-131.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement*, 4(4), 386-422.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189-227.
- Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (pp. 74-91). Cambridge, England: Cambridge University Press.
- Rinnert, C., & Kobayashi, H. (2001). Differing perceptions of EFL writing among readers in Japan. *The Modern Language Journal*, 85(2), 189-209.
- Ross, S. (1997). An introspective approach to understanding inference in a second language listening test. In G. Kasper (Ed.), *Communication strategies: Psycholinguistic and sociolinguistic perspectives* (pp. 216-237). London, England: Longman.
- Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing*, 25(4), 553-581.
- Scales, J., Wennerstrom, A., Richard, D., & Wu, S. H. (2006). Language learners' perceptions of accent. *TESOL Quarterly*, 40(4), 715-738.
- Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465-493.
- Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, 18(3), 303-325.
- Smith, L. E., & Bisazza, J. A. (1982). The comprehensibility of three varieties of English for college students in seven countries. *Language Learning*, 32(2), 259-269.
- Smith, L. E., & Rafiqzad, K. (1979). English for cross-cultural communication: The question of intelligibility. *TESOL Quarterly*, 13(3), 371-380.

- Stone, M., & Wright, B. D. (1988). *Separation statistics in Rasch measurement* (Research Memorandum No. 51). Chicago, IL: MESA Press.
- Tennant, A., & Pallant, J. F. (2008). Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 20(1), 1048-1051.
- Tauroza, S., & Luk, J. (1997). Accent and second language listening comprehension. *RELC Journal*, 28(1), 54-71.
- Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex.
- Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11(2), 197-223.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15(2), 263-287.
- Weil, S. A. (2001). Foreign-accented speech: Encoding and generalization. *Journal of the Acoustical Society of America*, 109, 2473 (A).
- Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1-2), 28-35.
- Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3), 305-335.
- Wigglesworth, G. (1994). Patterns of rater behaviour in the assessment of an oral interaction test. *Australian Review of Applied Linguistics*, 17(2), 77-103.
- Wolfe, E. W., & Dobria, L. (2008). Applications of the multifaceted Rasch model. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 71-85). Thousand Oaks, CA: Sage.
- Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, 15(4), 465-492.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 8(3), 370.
- Wright, B. D., & Masters, G. N. (2002). Number of person or item strata $(4G+1)/3$. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 16(3), 888.

Xi, X., & Mollaun, P. (2009). *How do raters from India perform in scoring the TOEFL iBT speaking section and what kind of training helps?* (TOEFL iBT Research Report No. RR-09-31). Princeton, NJ: ETS.

Notes

- ¹ Researchers (Dewey, 2004; Malabonga, Kenyon, & Carpenter, 2005) have found that self-assessment can be an effective and reliable measure of language proficiency in certain contexts. A more formal and rigorous assessment would outweigh the purposes of the task. In our case, the self-assessments provided a way to measure the raters' proficiencies on a potentially large sampling of languages and allowed us to obtain a measure of proficiency for raters concerning some languages that we did not have the means, funds, or time to directly test.
- ² Note that we do not know whether the ETS raters assigned each individual test taker identical speaking scores across all six tasks, or whether each test taker's scores showed some variation across the six tasks—the data we received from ETS regarding the test takers' speaking test scores were averages across all six tasks.
- ³ A copy of the conversion table that ETS provided us may be obtained by e-mailing the authors.
- ⁴ *Statistically distinct* levels are defined as those that are three standard errors apart, centered on the mean of the sample. They are three standard errors apart “because this is conveniently more than $1.96 * \text{sqrt}(2) = 2.77$ SE, the distance corresponding to .05 significance” (Wright & Masters, 2002, p. 888). As Fisher (1992) explained, “The functional range of [a set of] measures is around 4 True SD [the standard deviation of reported measures corrected for measurement error inflation]. Inflate this by 1 RMSE [root mean square error, the average measurement error of reported measures] to allow for the error in the observed measures. Set a significance difference between two measures at 3 RMSE . Then there are $(4 \text{ True SD} + \text{RMSE}) / (3 \text{ RMSE}) = (4G+1)/3$ significantly different levels of measures in the functional range” (p. 238). For detailed explanations of this formula and its application, see Fisher (1992) and Wright and Masters (2002). The most recent versions of FACETS report the separation index in addition to the separation ratio, so users of FACETS no longer have to calculate the separation index by hand.
- ⁵ FACETS uses Welch's (1947) refinement of the Student's *t* test for possibly unequal variances when comparing the means of two samples (or estimates). See Linacre (2010, p. 262-263) for details of the actual statistical tests employed.

Appendix A
Background Questionnaire

Participant ID # _____ (To be filled in by the researcher)

ETS/TOEFL COE RESEARCH PROJECT

***PLEASE FILL OUT THE FOLLOWING BACKGROUND INFORMATION.
PLEASE PRINT CLEARLY.***

1. Name:

a. First name: _____

c. Middle initial: _____

b. Last name: _____

2. Age: _____

3. Gender: Male Female

4. Phone number: () _____ - _____

5. Email address: _____

(This will be your username for the computer program.)

6. Address (for mailing your payment):

Street

Apt. #

City State Zip

7. Social security number (for processing your payment): _____ - _____ - _____

8. Native language (first fluent language, also known as your “mother tongue”):

a. If your native language is other than English, how did you learn English?

b. If your native language is other than English how old were you when you started learning English?

9. Language you speak at home: _____

10. What languages, other than English, do you speak or have you studied or are currently studying? Please report and answer questions for each language other than English that you speak or have studied or are currently studying.

LANGUAGE A.	HOW DID YOU LEARN THE LANGUAGE? (Please describe.)	From what age to what age did you learn the language? ___ to ___	HOW WELL DO YOU SPEAK THE LANGUAGE? (Please circle one) poor / fair / good / advanced/ fluent / native-like Comments:
LANGUAGE B.	HOW DID YOU LEARN THE LANGUAGE? (Please describe.)	From what age to what age did you learn the language? ___ to ___	HOW WELL DO YOU SPEAK THE LANGUAGE? (Please circle one) poor / fair / good / advanced/ fluent / native-like Comments:
LANGUAGE C.	HOW DID YOU LEARN THE LANGUAGE? (Please describe.)	From what age to what age did you learn the language? ___ to ___	HOW WELL DO YOU SPEAK THE LANGUAGE? (Please circle one) poor / fair / good / advanced/ fluent / native-like Comments:
LANGUAGE D.	HOW DID YOU LEARN THE LANGUAGE? (Please describe.)	From what age to what age did you learn the language? ___ to ___	HOW WELL DO YOU SPEAK THE LANGUAGE? (Please circle one) poor / fair / good / advanced/ fluent / native-like Comments:

11. Do you have friends or family who speak any of the languages you listed above (in #10)?

Yes No

If Yes, Please explain.

12. Have you lived in or traveled to a place where people speak the languages you speak or have studied or are currently studying (the ones listed in #10)?

Yes No

If **yes**, please report and answer questions for each place you have lived or visited and where the language(s) (#10) were spoken.

Where did you travel or live? a. _____	For how long were you there?	How old were you when you were there?	What was the purpose of your visit or stay?
Where did you travel or live? b. _____	For how long were you there?	How old were you when you were there?	What was the purpose of your visit or stay?
Where did you travel or live? c. _____	For how long were you there?	How old were you when you were there?	What was the purpose of your visit or stay?

13. Are you now or have you ever been an English as a Second or Foreign Language (ESL/EFL) teacher?

Yes No

a. If yes, for how long (total)?

1 year or less 2-5 years 5-10 years More than 10 years

b. If yes, what state(s) (US) or country (countries) did you teach in?

a. _____ How long did you teach there? _____

b. _____ How long did you teach there? _____

c. _____ How long did you teach there? _____

Appendix B

Sample Rating Sheet

Sample number _____ Rater number _____

Scoring standards					Notes
	General description	Delivery	Language use	Topic development	
4	The response fulfills the demands of the task, with at most, minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Generally well-placed flow (fluid expression). Speech is clear. It may include minor lapses, or minor difficulties with pronunciation or intonation patterns, which do not affect overall intelligibility	The response demonstrates effective use of grammar and vocabulary. It exhibits a fairly high degree of automaticity with good control of basic and complex structures (as appropriate). Some minor (or systematic) errors are noticeable but do not obscure meaning.	Response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear (or clear progression of ideas).	
3	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected).	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. This may affect overall fluency, but it does not seriously interfere with the communication of the message.	Response is mostly coherent and sustained and conveys relevant ideas/information. Overall, development is somewhat limited, usually lacks elaboration or specificity. At times, relationships between ideas may not be immediately clear.	
2	The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation or choppy rhythm/pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and/or general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitions. Connections of ideas may be unclear.	
1	The response is very limited in content and/or coherence or is only minimally connected to the task, or speech is largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation, stress, and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limits (or prevents) expression of ideas and connections among ideas. Some low-level responses may rely heavily on practiced or formulaic expressions.	Limited relevant content is expressed. The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete task and may rely heavily on repetition of the prompt.	

