Brigham Young University
BYU ScholarsArchive

All Theses and Dissertations

2017-03-01

Investigating the effects of Rater's Second Language Learning Background and Familiarity with Test-Taker's First Language on Speaking Test Scores

Ksenia Zhao
Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd

 Part of the Linguistics Commons

BYU ScholarsArchive Citation

Zhao, Ksenia, "Investigating the effects of Rater's Second Language Learning Background and Familiarity with Test-Taker's First Language on Speaking Test Scores" (2017). *All Theses and Dissertations*. 6256.
https://scholarsarchive.byu.edu/etd/6256

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu.

Investigating the Effects of Rater's Second Language Learning Background and Familiarity with Test-Taker's First Language on Speaking Test Scores

Ksenia Zhao

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Troy L. Cox, Chair
Dan P. Dewey
Grant Eckstein

Department of Linguistics and English Language

Brigham Young University

Copyright © 2017 Ksenia Zhao

All Rights Reserved

ABSTRACT

Investigating the Effects of Rater's Second Language Learning Background and Familiarity with Test-Taker's First Language on Speaking Test Scores

Ksenia Zhao
Department of Linguistics and English Language, BYU
Master of Arts

Prior studies suggest that raters' familiarity with test-takers' first language (L1) can be a potential source of bias in rating speaking tests. However, there is still no consensus between researchers on how and to what extent that familiarity affects the scores. This study investigates raters' performance and focuses on not only how raters' second language (L2) proficiency level interacts with examinees' L1, but also if raters' teaching experience has any effect on the scores. Speaking samples of 58 ESL learners with L1s of Spanish (n = 30) and three Asian languages (Korean, n = 12; Chinese, n = 8; and Japanese, n = 8) of different levels of proficiency were rated by 16 trained raters with varying levels of Spanish proficiency (Novice to Advanced) and different degrees of teaching experience (between one and over 10 semesters). The ratings were analyzed using Many-Facet Rasch Measurement (MFRM). The results suggest that extensive rater training can be quite effective: there was no significant effect of either raters' familiarity with examinees' L1, or raters' teaching experience on the scores. However, even after training, the raters still exhibited different degrees of leniency/severity. Therefore, the main conclusion of this study is that even trained raters may consistently rate differently. The recommendation is to (a) have further rater training and calibration; and/or (b) use MFRM with fair average to compensate for the variance.

Keywords: language testing, rater bias, speaking tests, oral proficiency, language learning background, accented speech

ACKNOWLEDGEMENTS

First and foremost I would like to sincerely thank my graduate committee chair, Dr. Troy Cox, for his invaluable advice, expertise, guidance, and patience during the research design process, data collection and analysis, as well as the multiple revisions. Without his help this research project would not have taken place. I would also like to thank the other two members of my graduate committee: Dr. Dan Dewey and Dr. Grant Eckstein for their feedback and insight about research design and manuscript revisions.

I am grateful to my loving husband for the support and motivation he gave me on the way. The completion of this project would have taken much longer without his support and understanding.

I would also like to thank my dear family and friends for the encouragement and comfort they gave me when I needed it.

PREFACE

In accordance with the TESOL MA program guidelines, this thesis was written to be submitted to the journal Language Testing. This journal was chosen because its profile and audience are suitable for this study: it facilitates the exchange of ideas pertaining to language testing and assessment internationally. Moreover, this thesis builds upon a study of Winke, Gass, and Myford, (2012) previously published in the same journal. In their article titled "Raters' L2 background as a potential source of bias in rating oral performance" Winke et al. attempted to determine whether familiarity with examinees' L1 leads to rater's bias by examining ratings of 107 raters (trained online) scoring 432 TOEFL iBT speech samples from 72 test takers. The scores Winke et al examined were assigned by raters who were L2 learners/speakers of Spanish, Korean, and Chinese. The 432 samples were elicited from native speakers of Spanish, Korean, and Chinese respectively. Using the MFRM analysis, they concluded that L2 Spanish and Chinese raters were considerably more lenient to L1 Spanish and Chinese test-takers respectively. However, Korean L2 speaking raters did not exhibit any significant variability. The current study involves 16 trained raters with different levels of Spanish proficiency assessing speaking samples of 58 ESL students of the same three L1 backgrounds (with the addition of Japanese) as in the study by Winke et al. Moreover, this study also examines raters' varying lengths of teaching experience as a possible source of bias.

The articles for submission to Language Testing should follow the American Psychological Association's guidelines and be between 4000 and 8000 words, however this thesis currently contains four chapters and is 10256 words long, with the understanding that before submission some sections will be removed and made available as online supplemental material.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Speaking assessment has been a subject of debate among researchers for decades, due to its complexity resulting from the many possible sources of unwanted variance involved in rater-mediated assessment. Unlike selected response tests, which do not involve subjective scoring, speaking tests require raters to form judgments while assessing examinees' performance.

Examinees' performance irrelevant factors which may influence the outcome of a test are often referred to as rater bias or rater effects. As defined by Scullen, Mount and Goff (2000), rater effects are a "broad category of effects [which refers to] the systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the ratee" (p. 957). Huang (2013), describes rater effects as "construct-irrelevant variation associated with rater characteristics which are thus critical to the reliability, validity, and fairness of the speaking assessment" (p. 770). To ensure adequate assessment, speaking tests require raters who are able to refrain from any performance-irrelevant judgments.

Certainly, there are many factors which may affect raters' judgments in speaking tests such as rater's native (L1) or second language (L2) learning experiences, teaching experience or even previous personal encounters with a certain language or ethnicity, etc. Although there are studies which analyze one or several of the factors mentioned above, the nature of rater decision-making processes involved in subjective performance tests is too complex to warrant any firm conclusions. Of interest to this study is whether the interaction between raters' L2 and examinees' L1 affects speaking test scores and how that effect can be moderated by raters' teaching experience. While there are prior studies which suggest that raters' familiarity with test-takers' first language could be a potential source of bias, most of them focus on raters' accent familiarity or language learning experience as factors affecting raters' performance without

investigating how the level of raters' L2 proficiency can affect their judgment. This study investigates if the level of raters' proficiency in examinees L1 could also affect the scores. Additionally, this study examines the possibility of the length of rater's teaching experience as another potential source of bias.

1.1 Background

The phenomenon of rater bias is not caused by what an examinee says, or how he or she says it. But rather, bias is mainly associated with the rater's reactions and cognitive processes involved in decision making and actual rating. When rating an oral response, conducted in person or recorded, a rater has to engage in a complex, cognitively demanding process which includes "observing, recalling information from memory storage… organizing, combining, weighing, and integrating that information to draw inferences about individuals" (Myford & Wolfe, 2003, p. 387). However, when drawing these inferences, a rater has to refrain from any concepts, ideas or stereotypes, as well as preferences and prejudices irrelevant to the examinee's language capabilities being assessed. This can be difficult, considering firstly the diversity of potential examinees' backgrounds, languages, accents and behavioral patterns, and, secondly, raters' linguistic background, prior language or accent exposure and language learning or teaching experience.

1.1.1 Rater's L2 as a source of bias. Prior research on the issue of rater bias suggests that rater's familiarity with examinee's L1 can be a potential source of bias (Carey, Mannnell, & Dunn, 2011; Huang, 2013; Winke & Gass 2013; Winke et al, 2012; Xi & Mollaun, 2009). The degree of such familiarity may be different: from mere accent recognition to high proficiency in examinees' L1s. Some accents may be perceived by certain raters with various degrees of effort. According to Winke et al. (2012), "familiarity with a particular accent makes that type of

accented speech easier to understand than speech with an unfamiliar accent" (p. 232) because of adaptation. This kind of adaptation differs in levels: from basic knowledge about the language to prolonged study or living in the particular language environment to bilingualism. In each particular case the level of familiarity/adaptation to a certain accent/L1 may serve as a potential source of rater's leniency or severity to a specifically accented speech.

Processing and assessing accented speech, however, does not necessarily imply difficulties in comprehension. Trofimovich and Isaacs (2012) attempted to determine which linguistic aspects of second language speech related to accent and which to comprehensibility. They examined the analysis of 40 speech samples of native French speakers of English conducted by 60 novice raters and three experienced teachers. Their results found that although the two concepts overlap, accent is strongly correlated with the aspects of phonology, whereas comprehensibility is mainly influenced by grammatical and lexical accuracy. These findings reaffirm that even heavily-accented speech may still be completely comprehensible, grammatically accurate, and lexically rich, and that an accent should not have a negative impact on scores unless it impedes understanding.

Although previous research on the issue of accent familiarity or rater's L2 learning background as a potential source of bias has been conducted (Carey et al. 2011; Huang, 2013; Winke & Gass, 2013; Winke et al, 2012; Xi & Mollaun, 2009), there is no consensus among the researchers as to how rater familiarity with a test-takers' L1 actually affects scores. For instance, Huang (2013) investigated the impact of accent familiarity and ESL/EFL teaching experience on raters' performance. The results showed that there were no significant effects of accent familiarity or prior teaching experience (although, raters who were familiar with the accent and possessed prior teaching experience self-reported that their experiences did affect their ratings).

On the other hand, the results of the research conducted by Carey et al. (2011) suggested the opposite. Their experiment involved 99 trained raters living and working in different countries: Australia (n = 19), Hong Kong (n = 20) India (n = 20), Korea (n = 19), and New Zealand (n = 21) who rated three oral interviews with Chinese, Korean, and Indian-accented speech. The results indicated that familiarity with examinees accents/L1s had an effect on the ratings: raters with more exposure to certain L1s assigned higher ratings. The location of raters also had an impact: examinees who were scored in their home countries tended to receive higher ratings.

Another study by Winke et al. (2012) was conducted to determine whether familiarity with examinees' L1 leads to rater's bias by examining ratings of 107 raters (trained online) scoring 432 TOEFL iBT speech samples from 72 test takers. The scores Winke et al examined were assigned by raters who were L2 learners/speakers of Spanish, Korean, and Chinese. The 432 samples were elicited from native speakers of Spanish, Korean, and Chinese respectively. Utilizing MFRM (which is typically used for simultaneous analysis of multiple variables potentially having impact on assessment outcomes), they concluded that L2 Spanish and Chinese raters were considerably more lenient to L1 Spanish and Chinese test-takers respectively. However, Korean L2 speaking raters did not exhibit any significant variability.

It is worth mentioning that the first study mentioned above (by Huang in 2013) used untrained raters, whereas the other two involved trained raters: Carey et al (2011) used professional IELTS raters, while the raters in the study by Winke et al (2012) received a standard ETS online training for new raters.

Evidently, it is unclear as to what circumstances and to what extent rater's familiarity with test-takers L1/accent affects oral proficiency assessment results. Moreover, the question of

how the degree of rater's proficiency in test-takers L1 (or level of exposure to it) affects raters' performance also remains open with Winke et al. (2012) suggesting further exploration of this topic.

1.1.2 Rater's teaching experience as a source of bias. The other important aspect of the issue to take into consideration is raters' prior language teaching experience. Previous studies on the effect of teaching experience on speaking tests' ratings have yielded mixed results. For instance, while there are studies which suggest teachers are harsher raters than non-teachers (Hadden, 1991), other studies (Hsieh, 2011) indicate the opposite. Finally there is a body of research which suggests that there is no significant difference in how lenient/severe teachers and non-teachers are (Huang, 2013). Despite these mixed results, studies which involved different rating criteria tend to agree that even though teachers and non-teachers did not vary significantly in their overall ratings, they tended to approach rating process differently and prioritize different assessment categories (Hsieh, 2011; Huang, 2013).

Bailey (2004) suggests that in ESL/ EFL class settings "teachers react to mistakes more often than non-teachers do. In addition, non-native speaking teachers react more often than native-speaking teachers suggesting that teachers – especially non-native teachers – may be more sensitive to errors than other people" (p. 172). Undoubtedly, speaking assessments conducted by a teacher in an ESL/EFL class are different from one performed by a language proficiency test rater; however, the tendency of people with teaching experience to respond differently or, "more severely" (Bailey, 2004) to test-takers' inaccuracies should not be disregarded. Moreover, Hadden (1991) argues that in her study, which only involved NSs of English, the 25 teachers rated video samples of ESL students considerably more harshly than the 32 non-teachers.

On the other hand, a study by Hsieh (2011), which involved oral interviews of international teaching assistants, found that although overall there was no difference in scores between teachers and non-teachers, the non-teachers were noticeably harsher in their ratings on accentedness and comprehensibility.

Huang's (2013) research provided valuable insights into how untrained raters (with and without teaching experience and exposure to test takers' L1) treat Chinese accented speech. The study included three groups of raters who were asked to rate samples of 26 TOEFL iBT test takers. The raters were grouped according to (a) their familiarity with test takers' non-native accent (familiar or non-familiar with Chinese-accented English speech) and (b) English teaching background:

- 22 Unfamiliar Non-Teachers (UNTs)
- 22 Familiar Non-Teachers (FNTs) and
- 22 Familiar Teachers (FTs).

The analysis of the data revealed that the three groups did not exhibit significant differences in inter-rater reliability. Contrary to the main findings of the study, many raters self-reported that familiarity and previous teaching experience affected their decisions in the process of rating. Many raters believed that accent familiarity and TESL/TEFL experience enhanced their listening comprehension and error detection. There were also certain tendencies in regards to how different groups rated: (a) in rating overall proficiency UNTs were slightly more lenient than FNTs and FTs; (b) in the foreign accent dimension, however, FTs on average provided higher ratings than UNTs and FNTs; (c) as far as the grammar and vocabulary dimension was concerned, all three groups rated similarly; and finally, (d) in the content dimension, FNTs and FTs were slightly more lenient than UNTs. The involvement of untrained raters only, precludes

us from generalizing these findings onto trained raters. Moreover, none of the previously mentioned studies investigate whether trained raters rate differently depending on the length of their teaching experience. Hence, further research on the issue is needed.

1.2 Research Questions

1) To what extent are examinees' speaking test scores affected by the raters' L2 learning background?

2) To what extent are examinees' speaking test scores affected by the raters' teaching experience?

2. Methodology

The purpose of this study was to determine whether the degree of a rater's proficiency in an examinee's L1 and a rater's teaching experience should be considered as a source of bias. The study involved experienced raters working in an Intensive English Program (IEP) at a large university in Midwest with students having a variety of native languages. The most commonly spoken languages were Spanish, Portuguese, Korean, Chinese, Japanese, Russian, and French. For the purposes of this study, the following four of the most common languages were chosen: Spanish, Korean, Chinese, and Japanese. Since the most common raters' L2 and the most common examinees' L1 was Spanish, this study focused on the effect of possible interaction of Spanish as raters' L2 and Spanish as examinees' L1 on the test scores. The decision to focus on Spanish was largely determined by the specifics of the student and ESL teacher population which represents that of the American Midwest. While the focus on Spanish may limit the generalizability of the results to other languages, it allowed (a) consistent data analysis, and (b) continuation of the inquiry made by Winke et al. (2012), but with more rigorously trained raters.

2.1 Procedure

The procedure (which will be discussed more in depth in the subsequent sections) involved the following steps (see Figure 1): (a) prospective participants (raters) were invited to take a screening background survey (the data, collected in the survey was used for participants' selection and subsequent data analysis); (b) since the study used linked incomplete design, the selected raters were then divided into two groups with two different rating schedules; (c) each group rated 45 speaking samples (out of 58 selected from a database); and (d) a many-facet RASCH analysis (also referred to as many-facet Rasch measurement – MFRM) was performed on the ratings to determine if raters' L1 background and/or teaching experience caused any bias.

The detailed description of the participants and the instruments used in the study is provided in the subsequent sections.

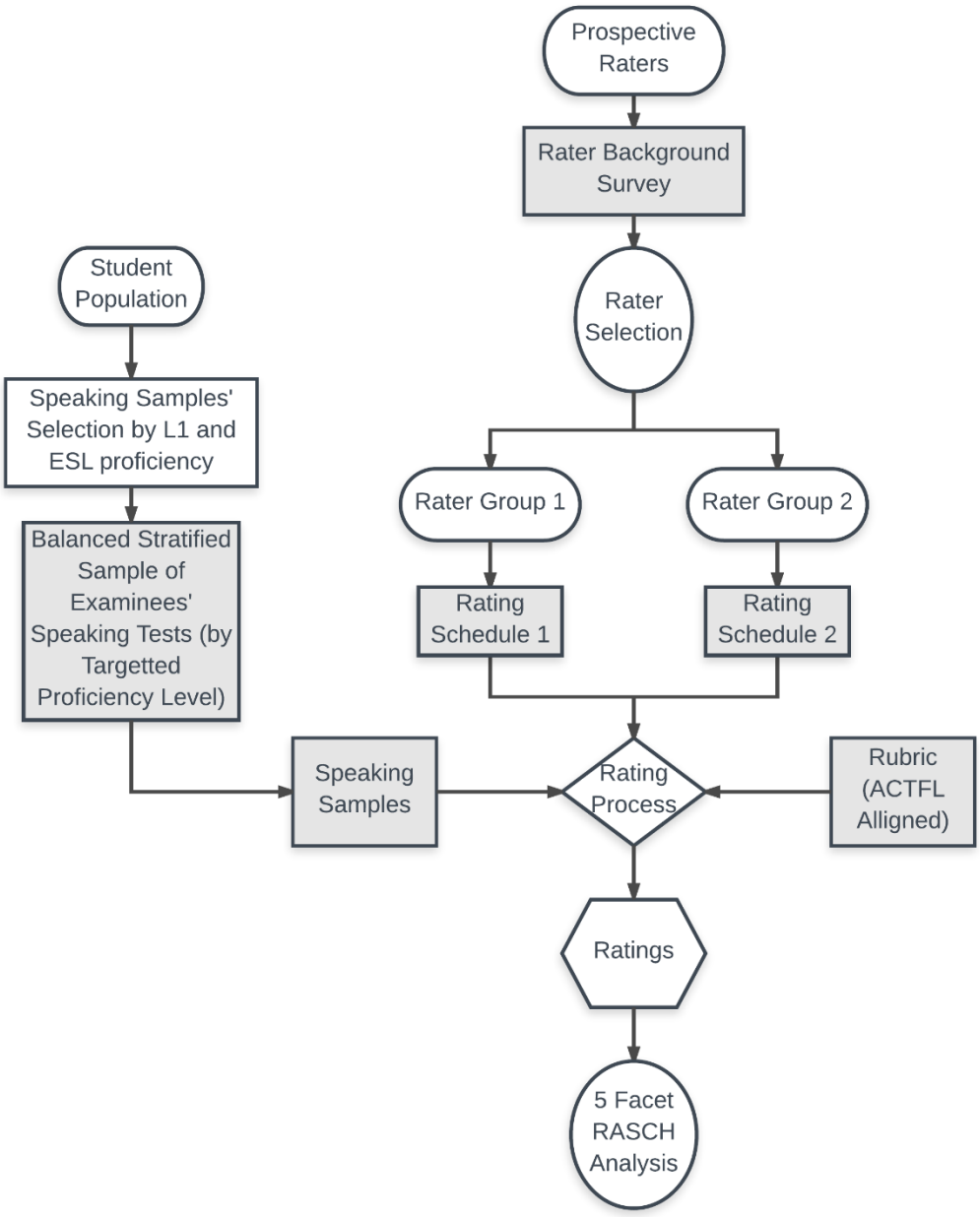

Figure 1. Study procedure.

2.2 Participants

Originally, 35 prospective raters took the screening background survey. The criteria for participation in the study required raters to be: (a) trained for the local major university's IEP; (b)

be a native speaker of English (to control for the native language variable as a source of variance); and (c) to be available during the experiment. Twenty two raters were willing to take part, but only sixteen raters met the criteria.

To avoid the Hawthorne effect (when the performance of subjects can be affected by their knowledge that they are being observed), the raters were told that they were a part of the IEP's study which evaluated the effect of using a reduced subset of questions from IEP's standard speaking assessment.

All participating raters had received prior rater training in the IEP where they were or had been employed as teachers. The ages of the participants ranged from early twenties to sixties (average = 32.3). Among the 16 participating raters, 14 were female and two were male. All of the participants were native speakers of English and spoke *at least* one language other than their native (Arabic [n = 1], Chinese [n = 1], Czech [n = 1], French [n = 4], German [n = 3], Hebrew [n = 2], Italian [n = 1], Japanese [n = 2], Korean [n = 2], Norwegian [n = 1], Portuguese [n = 2], Spanish [n = 12], and Turkish [n = 1]). Additionally, all of the participating raters had ESL teaching experience, ranging from one to 20 semesters in the university's IEP (some had taught at other schools as well). Two of the raters also had experience teaching languages other than English (Spanish).

Since the study used linked incomplete design involving two different rating schedules, the raters were randomly divided into two groups of eight, with the help of an online randomizer. Each group consisted of eight trained raters: group number one had six female and two male raters, while group number two had eight female raters.

Figure 2 shows the representation of different levels of Spanish ability and the length of teaching experience among the 16 participating raters. The IEP used American Council on the

Teaching of Foreign Languages (ACTFL) proficiency guidelines in rater and teacher training. Therefore, the ACTFL scale was chosen as the point of reference for language proficiency level: raters had to report their proficiency based on that scale (for more information on the ACTFL proficiency guidelines in speaking assessment see *ACTFL: Proficiency guidelines*). Based on their self-reported Spanish ability, the raters were divided into three groups: (a) *none* (or limited) Spanish ability (an equivalent of *Novice Low* to *Novice Mid* on the ACTFL scale) n = 6; (b) *mid* Spanish ability (*Novice High* to *Intermediate Mid* on the ACTFL scale) n = 7; and (c) *high* Spanish ability (*Intermediate High* to *Advanced Mid* on the ACTFL scale) n = 3. In terms of ESL teaching experience, participants were also divided into three groups: (a) *some* experience (raters who had taught for at least one to three semesters) n = 3; (b) *considerable* experience (raters who had taught for four to ten semesters) n = 6; and (c) *extensive* experience (raters who had taught for more than ten semesters) n = 7.

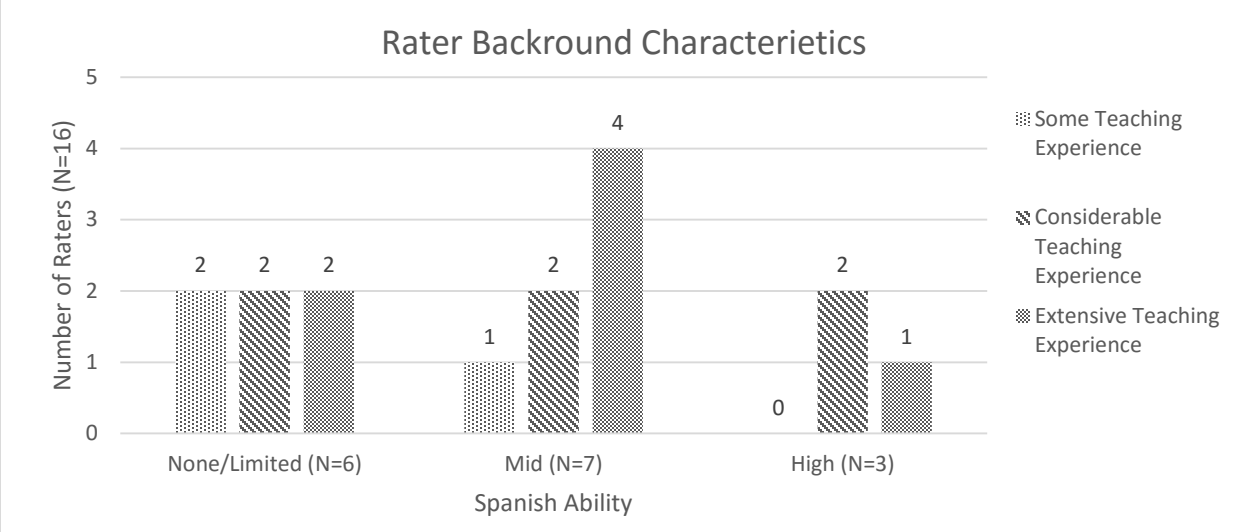

Figure 2. Rater Background Characteristics

2.3 Instruments

There were four instruments used in the study: rater background survey, rating schedule, speaking samples of ESL students, and an ACTFL-aligned English proficiency rubric.

2.3.1 Rater background survey. A screening background survey was administered to prospective participants online to gather demographic information: e.g. age, gender, languages spoken, circumstances of studying those languages, self-assessed level of proficiency, and teaching experience.

2.3.2 Rating schedule. The study used linked incomplete design which required the division of the 16 participating raters into the two groups of eight (See Figure 3). Out of the total number of 58 test-takers' samples, each group rated 45 samples: there was an overlap of 32 samples (or 71%) between the two rater groups. This overlap allowed us to have a sufficiently connected resulting dataset (Eckes, 2011). When creating the rating schedule the following factors were considered: (a) ensuring an equal, or the next best to equal, distribution of samples among the rater groups according to student gender within a certain L1 and ESL level subgroup; and (b) averaging the English language proficiency levels within each of the L1 subgroup among the two rater groups.

Sample No	2	3	4	5	1*	6	7	8	9	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#
Student L1	Chinese								Japanese								Korean											
Student Level	L	M	M	M	H	H	H	H	L	L	M	M	M	H	H	H	L	L	L	L	M	M	M	M	H	H	H	H
Student Gender	M	F	M	M	F	M	M	F	M	M	F	M	M	F	M	M	F	M	F	F	F	M	M	F	F	F	F	M
Rater group 1	X	X		X	X	X		X	X	X	X	X		X		X	X	X	X		X		X	X	X	X		X
Rater group 2	X	X	X		X		X	X	X	X	X		X	X	X		X	X		X	X	X		X	X		X	X

Sample No	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	#	TOTAL
Student L1	Spanish																														58
Student Level	L	L	L	L	L	L	L	L	L	L	M	M	M	M	M	M	M	M	M	M	H	H	H	H	H	H	H	H	H	H	
Student Gender	F	F	F	F	F	M	M	F	F	M	F	M	F	M	M	F	F	F	M	M	F	M	F	M	M	M	F	F	F	M	
Rater group 1		X	X	X	X	X	X	X		X	X	X	X	X			X	X	X	X	X	X	X		X	X	X		X	X	45
Rater group 2	X		X	X	X	X	X		X	X	X	X	X		X	X		X	X	X	X	X	X	X		X		X	X	X	45

Figure 3. Rating Schedule

2.3.3 Examinees and sample selection. This subsection describes how the speaking samples were selected and which parts of the original IEP's speaking test were chosen for the analysis.

2.3.3.1 Sample selection process. The study used audio recordings of final oral examination responces of IEP's ESL students (examinees). To meet the requirements of MFRM,

the experiment employed stratified random sampling based on (a) examinees' L1; (b) ESL proficiency (based on the scores previously assigned at the IEP to examinees in final oral examinations); and (c) gender (when possible). The sampling resulted in a total number of 58 English language learners' (ELL) sets of responses from the university's IEP. Among those, 30 were of L1 Spanish speakers, 12 of Korean, eight of Mandarin Chinese, and eight of Japanese. The ratio of the selected examinees of the four L1 backgrounds reflected that of the IEP student population. It was also assumed that having the speakers of the three Asian languages would provide a substantial contrast to Spanish speakers in terms of pronunciation.

Originally the following layout was planned: based on examinees' ESL proficiency, 18 samples were of *low* level, 20 were of *mid* level, and 20 of *high* level. However, because of a malfunction in the rating delivery system, one of the 58 samples was mislabeled. Original sample of examinee number 1 (selected as a L1 - Chinese, *low* ESL proficiency, male) was replaced with a sample of an L1 - Chinese, *high* ESL proficiency, female examinee. Therefore, the original layout slightly changed to the following: 17 samples were of *low* level, 20 were of *mid* level, and 21 of *high* level. Since the data analysis used MFRM (part of the Rasch family of Item Response Theory), which is both person and item independent, it is more important to have a full range of samples, rather than an equal distribution (Embertson & Reise, 2000).

The English proficiency levels were assigned according to the scores the examinees had previously received in the IEP end-of-semester examinations. The distribution of the L1s and English proficiency levels among the selected samples is shown in *Table 1* (for more detailed information on the examinees, see Appendix A). During the end-of-semester examinations the examinees were assessed by raters with the help of a rubric described in the next section (see Appendix C). Examinees who received MFRM fair averages ranging from 1.5 to 2.0 were

assigned *low* English proficiency level, examinees who received fair averages ranging from 3.0 to 3.5 were assigned *mid* proficiency level, and, finally, examinees who scored between 4.5 and 5.0 were assigned *high* proficiency level. When selecting the fair average ranges, an interval of 1.0 score difference between the groups was maintained to ensure that the examinees' groups were significantly different from each other in English proficiency. Because of the limited availability of the students of lower level in the database, there were two groups with a low number of examinees in them: *low* Chinese (n = 1) and *low* Japanese group (n = 2). When strictly adhering to the chosen score ranges was not possible (due to the unavailability of students of certain L2 background or English proficiency), minor deviations were allowed. There were 14 cases when the original scores of the students slightly deviated from the selected score ranges: the case with the afore-mentioned examinee number 1 with the score of 4.25; one case in the *low* Japanese group: a student with a score 2.01; three cases in the *high* Japanese group: students with scores 4.37, 4.44, 5.01; one case in the *low* Korean group: a student with a score of 2.01; and eight cases in the *low* Spanish group: 2.19, 2.2, 2.33, 2.44, 2.5, 2.61, 2.61, and 2.75. The cases in the *low* Japanese and *low* Korean groups were outside of the established range only by 0.01 and did not have any significant impact on the data distribution. The case with student number 1 who needed to be moved to the *high* Chinese group did not push the group average out of the desired range, although the student was 0.25 lower than the group threshold. The cases in the *high* Japanese and *low* Spanish groups caused the groups' averages to be out of the preferred group score range (4.6 in the *high* Japanese group and 2.3 in the *low* Spanish group), however, because of having an interval of 1.0 between the three major level groups, even with this deviation, the examinee groups were still sufficiently different from each other. This issue is also accounted for in the rating schedule where the examinees were distributed between the two rater groups in a

way that each rater group would have a comparable and representative sample with a significant overlap (71%) between the two groups.

Table 1

Distribution of Examinees According to Their L1 and Proficiency Level in English

		L1 background			
		Spanish	Korean	Japanese	Chinese
Group	Approximate ACTFL Equivalent				
Low	Novice Mid – Novice High	10	4	2	1
Mid	Intermediate Low – Intermediate Mid	10	4	3	3
High	Intermediate Mid Plus – Intermediate High	10	4	3	4
Total number of students (n = 58)		30	12	8	8

2.3.3.2 Speaking tests. Normally, on a speaking test, an ELL would need to answer 12 prompts of different levels of difficulty (*Novice* to *Superior* on ACTFL proficiency scale): one warm-up prompt, one *novice* prompt, three *intermediate* prompts, three *advanced* prompts, three *superior* prompts, and one cool-down prompt. For the purposes of this study, a subset of six questions of *novice* (n = 1), *intermediate* (n = 2), *advanced* (n = 2), and *superior* (n = 1) levels were selected. The selected prompts were targeted at language functions such as descriptions, past narration, and ability to ask and answer questions or support an opinion on an abstract idea. In the process of prompt selection it was ensured that the raters rated timed responses (in the form of audio recordings) of different test takers to the same six prompts (see Appendix B for a more detailed description of the prompts).

2.3.4 Rating rubric. When rating speaking samples, the participating raters referred to an ACTFL-aligned rating rubric used in the IEP for achievement tests and placement. Since the rubric had been used in multiple tests placement and achievement tests, it had been proved to be

a reliable rating instrument (Cox, 2013). The rubric encompassed eight levels of speaking proficiency (0 to 7) covering the three main assessment criteria: (a) text type (pertaining to fluency, discourse length and organization); (b) content; and (c) accuracy (see Appendix C).

2.4 Data Analysis

In order to answer the aforementioned research questions, the data was analyzed using Many-Facet Rasch Measurement (MFRM). FACETS (Linacre, 2015) was used to perform a 5 facet (variable) Rasch analysis. FACETS applies objective measurement principles from Rasch measurement to judged tests and can determine whether the analyzed variables are in any form of interaction, thus enabling one to see if there is any rater bias as a result of interaction between raters' and examinees' characteristics. The five facets used in the analysis were: 1 – examinees; 2 – raters; 3 – examinees' L1; 4 – raters' Spanish proficiency level; and 5 – raters' teaching experience. Below is a brief description of the five facets.

2.4.1 Facet 1: Examinees. The 58 examinees were coded for their (a) L1 (1 - Chinese, 2 - Japanese, 3 - Korean, 4 - Spanish); (b) ESL proficiency level (1 - *low*, 2 - *mid*, 3 - *high*); and (c) gender (1 - male, 2 - female). Although the gender variable was not included in this study, it was used in creating the rater schedule to make sure that both groups of raters receive an equal (where possible), or next best to equal, distribution of examinees of both genders.

2.4.2 Facet 2: Raters. The 16 raters were coded for their (a) Spanish proficiency level (1 - *none/limited*, 2 - *mid*, 3 - *high*); and (b) ESL teaching experience (1 - *some*, 2 - *considerable*, 3 - *extensive*).

2.4.3 Facets 3, 4, and 5: Dummy facets. Facets 3, 4, and 5 were dummy facets. A dummy facet is a kind of facet used for interaction analysis rather than for measuring main bias (see *Dummy facets for interactions*). Dummy facets often contain repetitive information about

the investigated traits and are helpful in interaction analysis. Facet 3 was coded for Examinees' L1: 1 - Chinese, 2 - Japanese, 3 - Korean, 4 - Spanish. Facet 4 was a coded for raters' Spanish ability level: 1 - *none/limited*, 2 - *mid*, 3 - *high*. Finally, facet 5 was a dummy facet coded for raters' length of ESL teaching experience: 1 - *some*, 2 - *considerable*, 3 - *extensive*.

3. Results

This section describes the results of the Many-Facet Rasch analysis performed on the data. First, we report the reliability of the instruments (the rubric and the samples selection criteria). After that, the FACETS variable map is discussed. Finally, we discuss the rater-related trends to answers the research questions.

3.1 Reliability of Instruments

3.1.1 Rubric. In order for the rubric to function adequately, the average examinee proficiency measures of the rating scale need to advance monotonically with each category (Eckes, 2011). According to *Table 2* (also visually represented in *Figure 4*), the average measures within each of the rubric's categories advanced monotonically from –6.10 to 5.13 on the logit scale. Another indicator of a rubric functioning adequately is the mean square outfit statistic (outfit MnSq): normally it should not exceed 2.0 (Eckes, 2011). According to *Table 2*, the outfit MnSq did not exceed 1.10 for any of the categories.

Table 2

Rubric Categories' Statistics

DATA					QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category	Counts		Cum.		Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
1	12	12	2%	2%	-6.10	-6.01	.9			(-8.16)		low	low	100%
2	73	73	10%	12%	-4.46	-4.41	.9	-7.04	.34	-5.71	-7.19	-7.04	-7.10	65%
3	186	186	26%	38%	-2.53	-2.43	.8	-4.38	.16	-2.92	-4.34	-4.38	-4.37	68%
4	241	241	34%	71%	.11	-.01	1.0	-1.50	.12	.02	-1.47	-1.50	-1.49	70%
5	156	156	22%	93%	2.15	2.17	1.0	1.56	.13	2.88	1.51	1.56	1.53	65%
6	47	47	7%	99%	3.83	3.83	1.1	4.22	.19	5.69	4.25	4.22	4.22	68%
7	4	4	1%	100%	5.13	5.65	1.1	7.15	.57	(8.26)	7.27	7.15	7.18	100%
										(Mean)		(Modal)	(Median)	

Additionally, *Figure 4* demonstrates probability curves of rubrics' categories. According to *Figure 4*, probability curves for each category are dispersed evenly and each category has a distinct peak. This suggests that the rubric used by raters functioned adequately (Eckes, 2011). Thus, this study concurs with findings in Cox (2013) that the IEP's rubric is validated and functions

properly for the purpose for which it was created, that is to differentiate between speakers with differing levels of English proficiency.

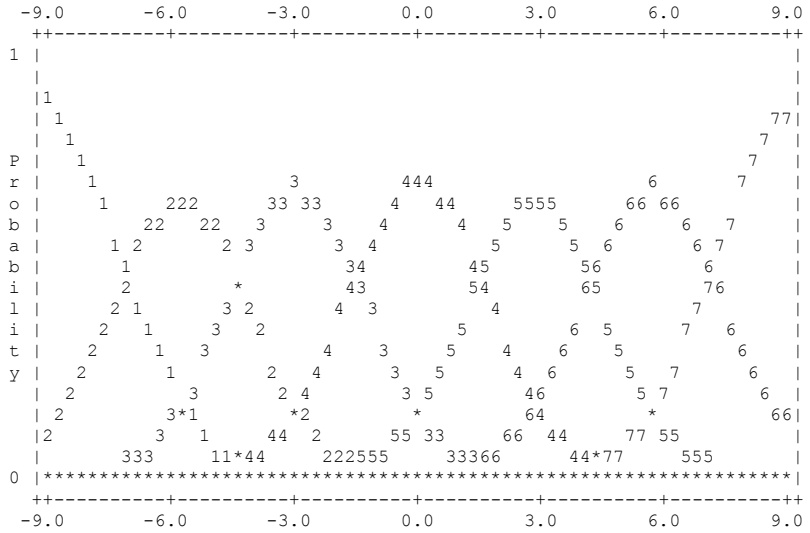

Figure 4. Rubric: Probability Curves

3.1.2 Examinees' samples selection criteria. The sample selection criteria were designed to have a range of ability levels. *Figure 5* shows a scatterplot of correlation between the original examinees' ratings from the database and the ratings acquired from the participating raters. With Pearson's *r* of .89 ($r^2 = .79$) between the original examinees' scores and the scores obtained in the rating process, all three levels of examinees' ESL proficiency form distinct well-defined groups. This suggests that the three ESL proficiency levels, originally chosen in the selection process, functioned as intended. According to this data, the selection of the six prompts targeting different ability levels had a strong correlation with examinees' performance on the tests containing the original set of 12 prompts.

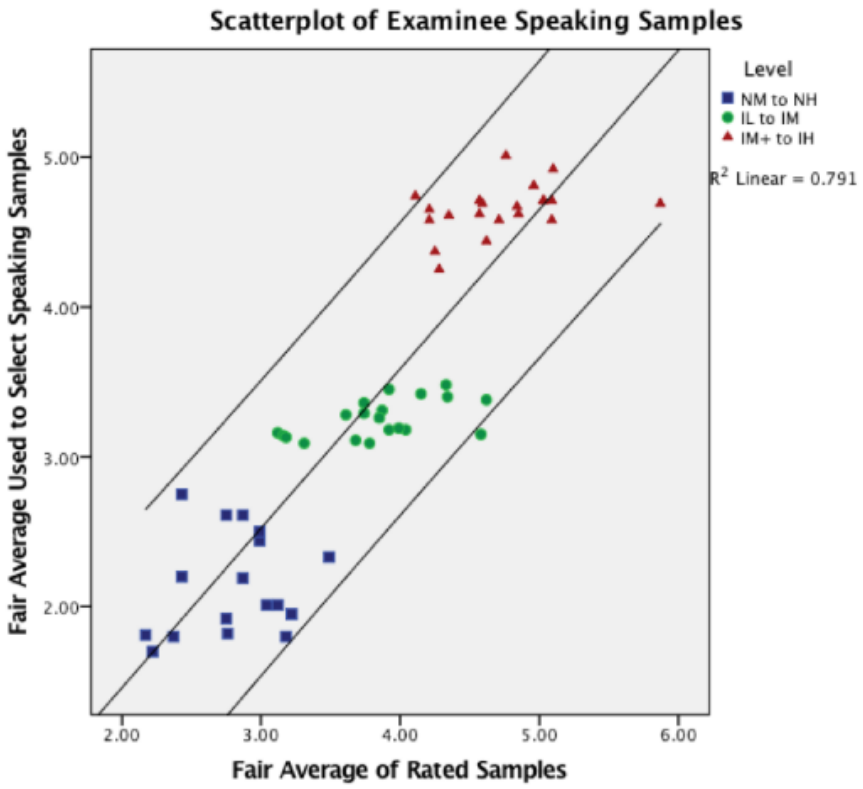

Figure 5 Scatterplot of Examinees Speaking Samples

Table 3 shows the examinees-related results of the data analysis. RASCH analysis suggests that examinees were reliably different form each other. The reliability of separation was .96, which indicates that there are statistically significant differences between examinees (Myford & Wolfe, 2003). The examinee separation ratio was 4.85 with stratification of 6.80 (see Table 3). These values indicate that the examinees were statistically different from one another and could be divided into six statistically distinct strata of examinees of different levels of ESL ability. This is also reflected in the variable map (see Figure 6): there is a distribution of examinees of all L1 backgrounds across the logit scale. The difference between the examinees is understandable and even desirable, since the examinees were selected intentionally to be of different levels of ESL proficiency even within each of the four L1 groups. However, it is noteworthy that two examinees (number 32: L1 - Spanish, *low* ESL proficiency; and number 50: L1 – Spanish, *high* ESL

proficiency) had outfit MnSq values of 2.42 and 2.0 respectively (normally this index is expected to be below 2.0) [see Appendix D for details]. These comparatively high values of outfit MnSq indicate that the ratings of the two above-mentioned students varied more than expected by the RASCH model (Eckes). Student number 32 had an original (database) rating of 2.2, and was awarded two 1's, six 2's, seven 3's, and one 4 during the rating process. Student number 50 had an original rating of 4.58 and was awarded two 3's, nine 4's, four 5's, and one 6. Both students were rated by all 16 raters. The quality of both samples was sufficient. Unfortunately, we have no way of knowing where the mistake occurred. It is possible that data entry errors during the rating process resulted in high outfit MnSq values. Despite the two misfitting students, the separation statistics still suggest that the students were sufficiently different from one another and that the level separation worked as intended.

Table 3

Examinees Measurement Report

Logit Mean	-0.41
Logit SD	2.53
Fair Ave Mean	3.86
Fair Ave SD	0.87
Outfit Mean	0.92
Outfit SD	0.42
Separation Ratio	4.85
Stratification	6.80
Separation Reliability	.96

3.2 MFRM Analysis Results

The FACETS variable map below demonstrates the general trends pertaining to raters and examinees (see Figure 6).

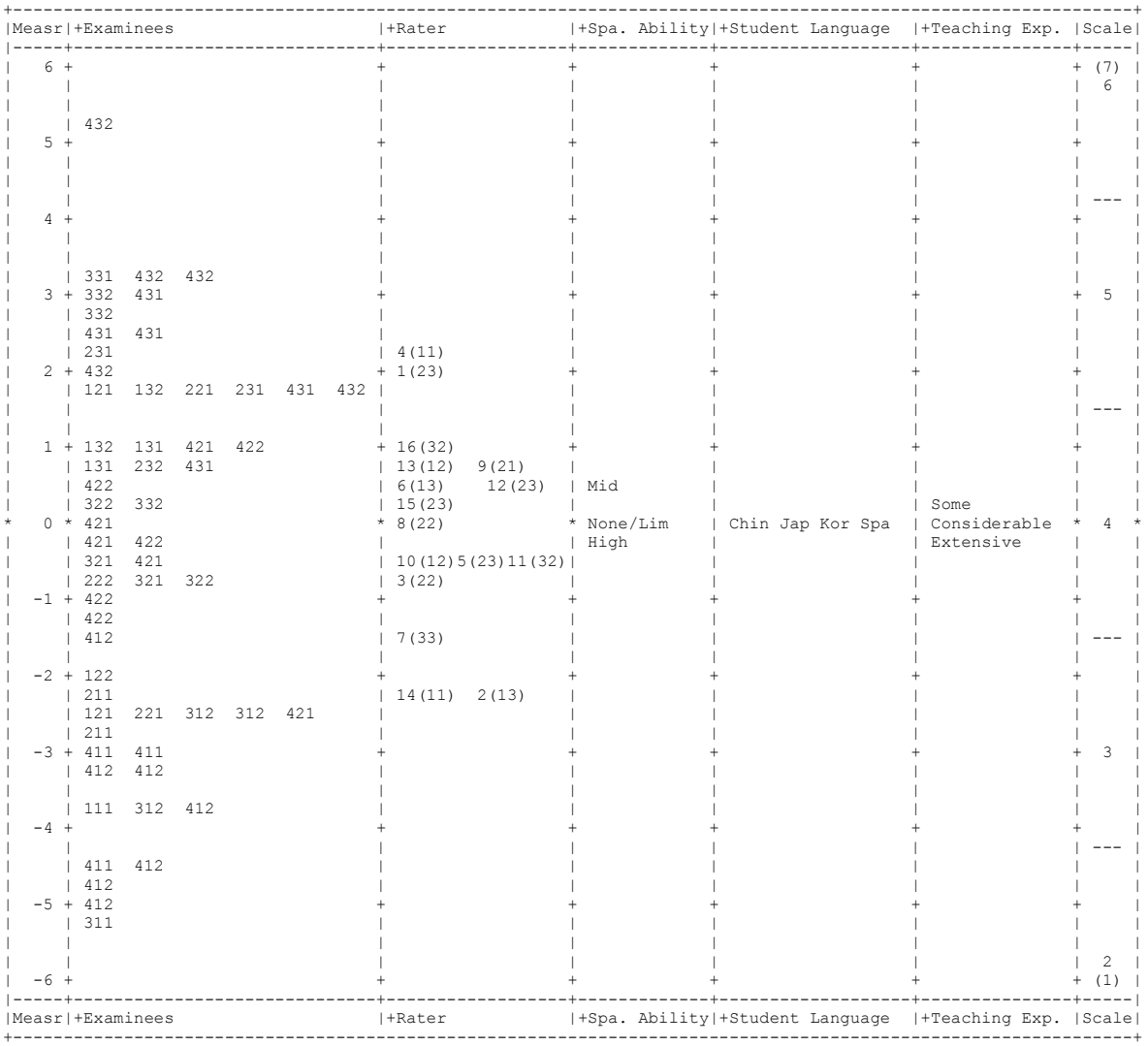

Figure 6. FACETS Output Variable Map

The first column of *Figure 6* shows a logit scale from -6 to +6 on which the measures of examinees and raters are positioned. A logit (short for "log odds unit") "…is a linear function of the ability and difficulty parameters. [W]hen the examinee ability equals the item difficulty the odd logs of success are zero" (Eckes, 2011). The negative end of the scale corresponds to the least able examinees and the least lenient raters.

The seventh column depicts the eight (0 to 7) scoring rubric categories (aligned to the logit scale) which raters used in the rating process. For instance, a score of 5 in the rubric corresponds to a 3 on the logit scale, and a score of 4 in the rubric corresponds to the zero on the logit scale.

The second column shows the positioning of examinees on the logit scale. The higher an examinee is on the logit scale, the higher ability he or she demonstrated. Each three-digit number represents one examinee: the first digit stands for L1: 1 - Chinese, 2 - Japanese, 3 - Korean, 4 - Spanish; the second digit stands for ESL proficiency level: 1 - *low*, 2 - *mid*, 3 - *high*; the third digit stands for gender: 1 - male, 2 -female. The most able examinee (coded 432) was at the top of the logit scale with the measure of 5.29 (which corresponds to a score of 6 in the rubric), while the least able examinee (coded 311) was at the bottom of the logit scale with the value of -5.26 (which corresponds to a rating of 2, according to the rubric).

The third column shows the positioning of raters on the logit scale. The higher a rater is on the scale, the more lenient he or she was, while rating the examinees. For each rater (numbered 1 to 16) there are coded characteristics in brackets: the first number stands for Spanish ability (1 - *none/limited*, 2 - *mid*, 3 - *high*), while the second number represents ESL teaching experience (1 - *some*, 2 - *considerable*, 3 - *extensive*). The raters are spread between the logit values of 2.32 and -2.32.

Columns 4 to 6 are the dummy facets in the analysis, used to examine interactions between the main facets (examinees and raters). The forth column shows how lenient or severe raters with different levels of proficiency in Spanish were to examinees. The higher a group is on the scale, the more lenient raters of that group were to examinees.The fifth column shows how examinees of the four L1 backgrounds were rated by different raters: the higher a particular L1 group is on the scale, the more lenient raters were to that particular group. The sixth column of *Figure 6* depicts how lenient or severe raters with different lengths of ESL teaching experience were to examinees. The higher a group is on the scale, the more lenient raters of that group were to examinees.

3.2.1 Raters. The reliability of the rater separation index (separation reliability) is generally preferred to be close to zero. When that is the case, the raters are believed to be interchangeable (Myford & Wolfe, 2003). The reliability of the rater separation index in this study was .96, which signifies that the raters differed in leniency/severity they exercised and were not interchangeable. Leniency/severity effect occurs when a rater or a group of them assigns higher or lower ratings than expected or warranted by some other performance criterion (Myford & Wolfe, 2003). *Table 4* suggests that raters were reliably different form each other: the rater separation ratio was 4.85 with stratification of 6.79. These values indicate that the raters were statistically different from one another and could be divided into six statistically distinct strata of leniency/severity. These values also mean that the "…differences between rater severities are over [four] times greater than the error with which these severities are measured" (Myford & Wolfe, 2003). Column 2 in the variable map (see Figure 6) demonstrates that most severe were raters number 14 and number 2, while the most lenient were raters number 4 and 1 (see details in Appendix E).

Table 4

Rater Measurement Report

Logit Mean	0.00
Logit SD	1.25
Fair Ave Mean	3.86
Fair Ave SD	0.42
Outfit Mean	0.95
Outfit SD	0.30
Separation Ratio	4.85
Stratification	6.79
Separation Reliability	.96

3.2.2 Research question 1: Levels of raters' Spanish ability as a source of bias. This subsection reports the results pertaining to the first research question of this study: to what extent

are examinees' speaking test scores affected by the raters' L2 learning background? If the raters' level of Spanish proficiency did not have effect on the ratings, the rater separation reliability would be close to zero. Although the variable map (Figure 6) shows that the raters were closely clustered near the zero on the logit scale according to their Spanish ability, the separation reliability of .82 indicates that there was a difference in how raters with different levels of Spanish ability rated the examinees. The separation ratio of 2.13 and the stratification of 3.17 (see Table 5) suggest that the raters can be divided into up to 3 different groups (according to their leniency/severity based on their Spanish ability). However, the difference was not as expected. Sinse Winke et al (2012) found that Spanish and Chinese L2 raters were more lenient with Spanish and Chinese L1 examanees, there was a hypothersis that the more familiar a rater was with examinees' L1, the more lenient he or she would be. This study did not see a progression based on the degree of familiarity. The raters with *mid* Spanish proficiency were 0.38 logits more generous to the examinees than the average of all the ratings. The *none/limited* and *high* Spanish speaking groups were more severe, with the *none/limited* at -.11 logits and the *high* at -.27 logits lower than the average of all the ratings. This finding is counter-intuitive. If this study followed the patterns of the other studies, those with more familiarity (*mid* and *high*) would be more generous than those with little or no familiarity. It is possible that this was caused by the small number (n = 3) of raters in the *high* Spanish ability group. The number of the raters of different Spanish ability groups may not have been large enough to identify how systematic the difference was.

Table 5

Raters' Spanish Proficiency Level Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Spanish Ability
485	135	3.59	3.77	-.27	.15	.95	-.4	.95	-.3	1.05	.83	.84	3 Adv
1002	269	3.72	3.82	-.11	.10	1.10	1.1	1.10	1.1	.90	.88	.87	1 Pre
1283	315	4.07	3.98	.38	.10	.81	-2.4	.81	-2.4	1.19	.85	.84	2 Mid
923.3	239.7	3.80	3.86	.00	.12	.95	-.6	.95	-.6		.86		Mean (Count: 3)
330.5	76.4	.20	.09	.28	.02	.12	1.5	.12	1.5		.02		S.D. (Population)
404.8	93.5	.25	.11	.34	.03	.14	1.8	.14	1.8		.03		S.D. (Sample)

Model, Populn: RMSE .12 Adj (True) S.D. .25 Separation 2.13 Strata 3.17 Reliability .82
Model, Sample: RMSE .12 Adj (True) S.D. .32 Separation 2.70 Strata 3.93 Reliability .88
Model, Fixed (all same) chi-square: 19.0 d.f.: 2 significance (probability): .00
Model, Random (normal) chi-square: 1.8 d.f.: 1 significance (probability): .18

3.2.2.1 Differences in the examinees' L1. If the examinees' L1 did not have effect on the ratings, the examinee separation reliability based on their L1s would be zero. The separation reliability of .00, separation ratio of 0.00, and stratification of 0.33 (see Table 6) suggest that examinees' L1s did not have effect on the ratings. This is also confirmed by the variable map (Figure 6): all the four examinee L1 groups are clustered close to each other with a range of .13 (mininmum = -.09 and maximum = .04) on the logit scale. Moreover, the large standard error values (e.g. Model S.E.—see Table 6) indicate that the groups were not statistically different. Thus, the data suggests that there was no statistically significant variation in rater severity based on examinees' L1: the students of the four different L1 backgrounds were treated fairly by different raters.

Table 6

Examinees' L1 Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Student Language
364	96	3.79	3.83	-.09	.17	.88	-.8	.88	-.8	1.12	.85	.81	2 Japanese
366	95	3.85	3.87	.02	.17	1.00	.0	1.00	.0	1.01	.83	.82	1 Chinese
555	144	3.85	3.87	.03	.14	.79	-1.8	.79	-1.8	1.21	.90	.87	3 Korean
1485	384	3.87	3.87	.04	.09	1.01	.1	1.01	.1	.99	.87	.87	4 Spanish
692.5	179.8	3.84	3.86	.00	.14	.92	-.6	.92	-.6		.86		Mean (Count: 4)
464.1	119.6	.03	.02	.05	.04	.09	.8	.09	.8		.02		S.D. (Population)
535.9	138.1	.03	.02	.06	.04	.10	.9	.11	.9		.03		S.D. (Sample)

Model, Populn: RMSE .15 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Sample: RMSE .15 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Fixed (all same) chi-square: .4 d.f.: 3 significance (probability): .94

3.2.2.2 Bias interactions. While there is no group level difference based on examinee L1, there is still the possibility of an interaction that might be confounding the results. Thus, following the methodology in Winke et al (2012), a MFRM bias analysis was performed. *Table 7* shows the summary statistics of each L1 group of examinees (Chinese, Japanese, Korean, and Spanish) rated by different Spanish ability groups of raters (*none/limited*, *mid*, *high*). The observed expected average in *Table 7* represents the difference between the observed and expected raw score divided by the observed count. *Table 7* also shows the bias size calculated in terms of raw score units. "The bias size is the size of the bias in logit units relative to the rater subgroup's overall severity measure" (Winke, 2012). Each bias size is also provided with a *t*-statistic, which is used with "degrees of freedom and the *p*-value to determine whether the interaction between subgroups was statistically significant" (Winke, 2012).

None of the interactions were statistically significant (only values with *p*-value less than .05 are considered to be statistically significant). Therefore, in regards to the first research question, there was no statistically significant interactions between examinees' L1s and raters' level of Spanish ability. These findings contrast with the results in Winke et al (2012), who did find statistically significant interactions between examinees' L1s and raters' L2s

Table 7

Bias Interactions

Bias/Interaction: 3. Student Language, 4. Spanish Ability (higher score = higher bias measure)

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Student Language Sq	N	Student	measr	Spanish Abil N	Spa	measr
174	171.69	42	.06	.16	.26	.60	41	.5506	.7	.7	5	1	Chinese	.02	2	Mid	.38
65	64.45	18	.03	.09	.40	.22	17	.8266	1.0	1.1	9	1	Chinese	.02	3	Adv	-.27
170	168.78	42	.03	.08	.26	.32	41	.7518	.8	.8	6	2	Japanese	-.09	2	Mid	.38
543	538.89	144	.03	.08	.14	.58	143	.5612	1.1	1.1	4	4	Spanish	.04	1	Pre	-.11
258	256.78	63	.02	.06	.21	.26	62	.7954	.7	.7	7	3	Korean	.03	2	Mid	.38
261	260.67	72	.00	.01	.20	.07	71	.9482	1.0	1.0	12	4	Spanish	.04	3	Adv	-.27
201	201.43	54	-.01	-.02	.23	-.10	53	.9209	.8	.8	3	3	Korean	.03	1	Pre	-.11
63	63.23	18	-.01	-.04	.40	-.09	17	.9283	.5	.5	10	2	Japanese	-.09	3	Adv	-.27
681	685.68	168	-.03	-.08	.13	-.61	167	.5418	.9	.9	8	4	Spanish	.04	2	Mid	.38
131	132.09	36	-.03	-.09	.28	-.31	35	.7597	1.1	1.1	2	2	Japanese	-.09	1	Pre	-.11
96	96.88	27	-.03	-.09	.33	-.29	26	.7763	.9	.9	11	3	Korean	.03	3	Adv	-.27
127	129.93	35	-.08	-.24	.29	-.84	34	.4081	1.3	1.3	1	1	Chinese	.02	1	Pre	-.11
230.8	230.88	59.9	.00	-.01	.26	-.02			.9	.9	Mean (Count: 12)						
183.5	183.76	46.0	.04	.10	.08	.43			.2	.2	S.D. (Population)						
191.7	191.93	48.0	.04	.11	.09	.45			.2	.2	S.D. (Sample)						

Fixed (all = 0) chi-square: 2.2 d.f.: 12 significance (probability): 1.00

3.2.3 Research question 2: Length of raters' teaching experience as a source of bias.

This subsection reports the results pertaining to the second research question of this study: to what extent are examinees' speaking test scores affected by the raters' teaching experience? If the length of raters' teaching experience did not have effect on the ratings, the rater separation reliability would be close to zero. The separation reliability by their teaching experience was .42 (see Table 8). The separation ratio of 0.85 and the stratification of 1.46 suggest that the raters could be hypothetically divided into one and a half of groups (according to their leniency/severity based on the length of teaching experience). Moreover, the variable map (Figure 6) shows that the raters were closely clustered near the zero on the logit scale according to the length of their teaching experience. Unlike the Spanish ability effect, the difference among raters based on the length of their teaching experience was systematic. Raters with *some* (one to three semesters) experience were 0.20 logits more generous to examinees; and raters with *considerable* experience (four to 10 semesters) did not exhibit any noticeable leniency/severity; finally, raters with *extensive* (more than 10 semesters) teaching experience were on average 0.27 logits more severe to the examinees (See table 9). The results on the length of teaching

experience as a source of bias appear to be systematic: the more experiences a rater has as a teacher, the more severe ratings he/she seems to assign, which concurs with Bailey (2004) and Hadden (1991).

Table 8

Raters' Teaching Experience Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N Teaching Experience
1197	314	3.81	3.81	-.17	.10	1.05	.6	1.06	.7	.95	.86	.86	3 Extensive
1032	270	3.82	3.85	-.03	.10	.78	-2.6	.77	-2.8	1.23	.87	.83	2 Considerable
541	135	4.01	3.93	.20	.15	1.02	.2	1.03	.3	.97	.88	.89	1 Some
923.3	239.7	3.88	3.86	.00	.12	.95	-.6	.95	-.6		.87		Mean (Count: 3)
278.6	76.2	.09	.05	.15	.02	.12	1.5	.13	1.6		.01		S.D. (Population)
341.2	93.3	.11	.06	.19	.03	.15	1.8	.16	1.9		.01		S.D. (Sample)

Model, Populn: RMSE .12 Adj (True) S.D. .10 Separation .85 Strata 1.46 Reliability .42
Model, Sample: RMSE .12 Adj (True) S.D. .15 Separation 1.26 Strata 2.01 Reliability .61
Model, Fixed (all same) chi-square: 4.6 d.f.: 2 significance (probability): .10
Model, Random (normal) chi-square: 1.4 d.f.: 1 significance (probability): .24

4. Conclusion and Discussion

Assessing speaking is undoubtedly a complicated issue because of many factors which may impact assessment results. Language assessment involving human raters is even more complicated because of all the rater-related factors which may interfere with the final scores.

While there is a body of research (Carey et al. 2011; Hadden, 1991; Hsieh, 2011; Winke & Gass, 2013; Winke et al, 2012; Xi & Mollaun, 2009) which suggests that raters are prone to different types of bias when assessing speaking, results seem to be inconclusive. Many conclude that raters are indeed prone to bias (Carey et al. 2011; Hadden, 1991; Hsieh, 2011; Winke et al, 2012; Xi & Mollaun, 2009). Some of those studies involved trained raters (Carey et al, 2011; Winke et al, 2012; Xi & Mollaun, 2009), while others used novice raters without prior experience or training (Hsieh, 2011; Huang, 2013). This study used only native English speaking raters trained at one language institution. The study included gathering information about raters L2 background and teaching experience to determine whether those two factors may result in bias. The main purpose of our inquiry was to determine whether rater training may help to eliminate or at least mitigate rater effects.

This study was conducted with the intention to build on the study published by Winke et al (2012). While Winke et al (2012) found that raters may exhibit bias when their L2 overlaps with examinees' L1, we did not reach similar conclusions. The results reported in the previous chapter lead us to the belief that with sufficient rater training and reliable assessment instruments (such as a rubric or a rating scale), rater effects can indeed be minimized to the point at which their impact is negligible.

Our findings are the following: (a) as intended, the examinees involved in our experiment were, indeed, statistically different from one another; (b) the raters were different in their

severity even though they were trained. In regards to the first research question of this study (i.e. to what extent are examinees' speaking test scores affected by the raters' L2 learning background?), the results suggest that raters' L2 learning background (different levels of Spanish ability) did not have a systematic or statistically significant effect on the ratings. In regards to the second research question (i.e. to what extent are examinees' speaking test scores affected by the raters' teaching experience?), we found that, while the length of raters' teaching experience may affect ratings, the degree of such effect is not crucial and can be accounted for with the help of MFRM. The effect or raters' teaching experience seemed to be systematic: the more experienced teachers are, the more severely they rate, which concurs with Bailey (2004) and Hadden (1991).

It should be noted that our results do not invalidate the research by Winke et al (2012). We believe that the reasons why our experiment revealed results different from Winke et al. (2012) are the following: (a) unlike Winke's study, ours involved more extensively trained raters; (b) while most of the raters participating in Winke's study were undergraduate students, we employed raters trained at one institution most of whom were graduate students.

The results of our experiment lead us to the three main conclusions: (a) extensive rater training and calibration can help raters focus on examinees' performance and refrain from performance-irrelevant judgments; (b) if a source of inter-rater variance (for example, raters' teaching experience) is established, MFRM can be used to compensate for it; and finally (c), a well-designed rubric/rating scale is necessary for raters to be able to assign fair scores. A well-designed and validated rubric helps raters to have a common understanding of the scale categories. Without a shared understanding of the scale categories, raters can interpret the categories differently and allow their personal impressions or other irrelevant factors to influence their judgment. Extensive rating training and calibration can help achieve adequate and

homogeneous understanding of the rating scale and each of its categories, minimizing the possibility of leniency/severity, central tendency, randomness, and other rater effects.

Our research was unique because under the same experiment we investigated not only how raters' L2 proficiency level (and its degree) would interact with examinees L1, but also if raters teaching experience would have any effect on the scores.

There are certainly some limitations to our study. First, while working with only native speakers of English may limit the generalizability of the results in the light of the increasing number of non-native raters in many languages worldwide, such a design allowed us to focus on the two potential sources of bias and, hence, have better control over the experiment by reducing the number of variables.

Second, the involvement of examinees of only the four L1 backgrounds (Chinese, Japanese, Korean, Spanish) could also limit the generalizability. However, the choice of languages was dictated by the student population and the decision to have a widely spoken language (such as Spanish) to be contrasted by languages with a significantly different phonological patterns (in this case Asian languages: Chinese, Japanese, and Korean). Additionally, the distribution of the examinees among language and level subgroups was not uniform due to the characteristics of the local student population. If the L1s of examinees were not the ones in which the raters had had exposure, then perhaps our findings would have replicated Winke et al's (2012) study.

Moreover, among the 16 participating raters, the distribution of them among the different subgroups in L2 proficiency and teaching experience was not uniform. However, the smallest number or raters within each corresponding subgroup was three, which still allowed us to have

sufficient amount of data to analyze. Thus, future studies may benefit from having an equal representation of participants among different subgroups, as well as a bigger sample.

Nevertheless, despite the limitations listed above, we believe that the results of this study are important and contribute well to the body of research on rater bias in speaking assessment.

The main recommendations of this study are: (a) ensuring adequate rater training which would cover an array of possible rater effects and their sources (raters' L1, L2, teaching experience, familiarity with examinees' culture, accent, or even prior personal experiences); (b) using reliable rating instruments (rubric) with clear non-overlapping categories and thresholds; (c) implementing rater calibration which would allow raters meet, rate, and discuss ratings together; (d) having multiple raters (when possible) to rate the same samples to ensure inter-rater reliability; and (e) using MFRM to compensate for possible variability: using fair average, rather than observed average.

Our suggestions for future research would include: conducting more studies examining the interactions between *multiple* possible sources of rater bias, such as raters' L2 backgrounds, their teaching experience, etc. Also, in order to confirm our findings and prove them more generalizable, future research involving well-balanced groups of raters and examinees based on their characteristics of interest is needed. And finally, studies on rater bias involving language students (as well as raters) of various language backgrounds would greatly contribute to the body of research on rater bias.

References

ACTFL: Proficiency guidelines. Available at: https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/english/speaking

Bailey, K., (2014) *Practical English Language Teaching: Speaking.* New York: McGraw-Hill Companies.

Carey, M. D., Mannel, R. H., & Dunn, P. K. (2011) Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28(2),* 201-219.

Cox, T. L. (2013) Investigating Prompt Difficulty in an Automatically Scored Speaking Performance Assessment (Doctoral dissertation, Brigham Young University). Retrieved from http://scholarsarchive.byu.edu/etd/3929/

Dummy facets for interactions. Retrieved from: http://www.winsteps.com/facetman/dummy.htm

Eckes, T. (2011) *Introduction to Many-Facet Rasch Measurement.* Frankfurt: Peter Lang.

Embertson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. New York: Psychology Press.

Hadden, B. L. (1991). Teacher and non-teacher perceptions of second-language communication. *Language Learning*, *41*, 1–24. doi: 10.1111/j.1467-1770.1991.tb00674.x

Hsieh, C.-N., 2011. Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency (Doctoral dissertation: Spain Fellow Working Papers in Second or Foreign Language Assessment 9). Retrieved from: https://etd.lib.msu.edu/islandora/object/etd%3A1282/datastream/OBJ/view

Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *System*, *41(3),* 770-785. Retrieved from http://www.sciencedirect.com/science/article/pii/S0346251X13000973

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.

Linacre, J. M. (2015) Facets computer program for many-facet Rasch measurement, version 3.71.4. Beaverton, Oregon: Winsteps.com.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4(4)*, 386-422.

Randomly assign subjects to treatment groups. Available at: http://www.graphpad.com/quickcalcs/randomize2

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85(6)*, 956-970.

Tauroza, S., Luk, J., (1997). Accent and second listening comprehension. *RELC Journal, 28(1),* 54-71.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15(4)*, 905-916. doi: 10.1017/S1366728912000168

Qu, Y., & Ricker-Pedley, K. L. (2013). Monitoring individual rater performance for the TOEIC® speaking and writing tests. Educational Testing Service. The Research Foundation for the TOEIC Tests: A Compendium of Studies, 2, 9.1-9.9. Retrieved from http://www.ets.org/research/policy_research_reports/publications/chapter/2013/jroi

Winke, P., &, Gass, S. (2013) The Influence of Second Language Experience and Accent Familiarity on Oral Proficiency Rating: A Qualitative Investigation. *Tesol Quarterly*, Vol. *47(4)*, 762-789.

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30(2)*, 231-252. doi: 10.1177/0265532212456968

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 RR-09-31). Princeton: Educational Testing Service.

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

Zhang, B., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher ratings: Competing or complementary constructs? *Language Testin*g, *28(1)*, 31-50.

Appendices

Appendix A: Examinees' Data

Sample number	L1	ESL Proficiency	Gender	Age	Score
2	Chinese	Low	M	33	1.92
		Desired range:	1.5-2.0	Subgroup average:	1.92
3	Chinese	Mid	F	38	3.09
4			M	23	3.15
5			M	25	3.16
		Desired range:	3.0 - 3.5	Subgroup average:	3.13
1	Chinese	High	F	27	4.25*
6			M	23	4.61
7			M	18	4.65
8			F	24	4.71
		Desired range:	4.5 – 5.0	Subgroup average:	4.555
9	Japanese	Low	M	19	1.95
10			M	28	2.01*
		Desired range:	1.5-2.0	Subgroup average:	1.98
11	Japanese	Mid	F	22	3.09
12			M	18	3.14
13			M	22	3.38
		Desired range:	3.0 - 3.5	Subgroup average:	3.20
14	Japanese	High	F	22	4.37*
15			M	21	4.44*
16			M	25	5.01*
		Desired range:	4.5 – 5.0	Subgroup average:	4.61
17	Korean	Low	F	23	1.8
18			M	18	1.81
19			F	31	1.82
20			F	22	2.01*
		Desired range:	1.5-2.0	Subgroup average:	1.86
21	Korean	Mid	F	21	3.18
22			M	24	3.29
23			M	24	3.31
24			F	22	3.36
		Desired range:	3.0 - 3.5	Subgroup average:	3.285
25	Korean	High	F	21	4.71
26			F	21	4.74
27			F	17	4.81
28			M	24	4.92
		Desired range:	4.5 – 5.0	Subgroup average:	4.795
29	Spanish	Low	F	21	1.7
30			F	24	1.8
31			F	24	2.19*
32			F	18	2.2*
33			F	24	2.33*
34			M	39	2.44*
35			M	26	2.5*
36			F	24	2.61*
37			F	17	2.61*
38			M	42	2.75*
		Desired range:	1.5-2.0	Subgroup average:	2.31
39	Spanish	Mid	F	29	3.11
40			M	18	3.13
41			F	28	3.18
42			M	46	3.19
43			M	18	3.26
44			F	18	3.28
45			F	29	3.4
46			F	23	3.42
47			M	21	3.45
48			M	22	3.48
		Desired range:	3.0 - 3.5	Subgroup average:	3.29
49	Spanish	High	F	17	4.58
50			M	21	4.58
51			F	21	4.58
52			M	24	4.62
53			M	40	4.62
54			M	29	4.67
55			F	42	4.69
56			F	24	4.69
57			F	23	4.71
58			M	26	4.71
		Desired range:	4.5 – 5.0	Subgroup average:	4.645

Numbers marked with * - values outside of the desired subgroup score range

Appendix B: Speaking Test Prompts

Number	Targeted level	Prompt function	Prompt	Response duration (sec)
1	Novice	Descriptions and ability to list items	Talk about what you are wearing today. List the clothes and as many things about them as you can.. Also talk about your reason for choosing to wear them.	30
2	Intermediate	Ability to ask questions	You want to plan a party for a neighbor who is moving next month. What are several questions you will ask your neighbor in order to plan a party that she would like?	30
3	Intermediate	Description	A friend who lives far away asks about what you do on the weekend. Describe your routine on most Saturdays from the morning to the evening. What do you do? Where do you go? Who are you with? How are your weekend activities different from other people you know that live in other places?	30-46
4	Advanced	Problem description and solution	You are working with a group of classmates to complete a group assignment. Your responsibility was to create a media presentation with information and pictures. On the day of the presentation, you lost the USB drive containing the presentation and all of the information the group had collected. Explain to your group members what happened and propose a series of actions that will make the situation better.	45-61
5	Advanced	Past narration	Retell a story from your life when you or someone you know won a prize or award. Include a detailed description of the events before, during and after this experience. How or why was this experience memorable to you?	45-61
6	Superior	Stating and supporting an opinion	Two friends are having a debate. One friend believes that playing video games is a waste of time and parents should prohibit their use. The other friend believes that children can acquire valuable skills from video games and parents should facilitate their use. Choose one side of this argument to support and explain your reasons for having your opinion.	80-92

Appendix C: Rubric

Level	ACTFL equivalent	Text Type	Content	Accuracy
Criteria		Fluency Development Organization	Functional Ability with the Language (Abstract vs. Concrete or Self-centric Language) Vocabulary	Grammar & Verb Tense Communication Strategies Pronunciation Native-like Comprehensibility
7	Advanced Mid and higher	Exemplified speaking on a paragraph level rather than isolated phrases or strings of sentences. Highly organized argument (transitions, conclusion, etc.). Speaker explains the outline of topic and follows it through.	Discusses some topics abstractly (areas of interest or specific field of study) Better with a variety of concrete topics Appropriate use of formal and informal language Appropriate use of a variety in academic and non-academic vocabulary	Grammar errors are extremely rare, if they occur at all; wide range of structures in all time frames Able to compensate for deficiencies by use of communicative strategies—paraphrasing, circumlocution, illustration—such that deficiencies are unnoticeable Intonation resembles native-speaker patterns; pronunciation rarely if ever causes comprehension problems Pausing and redundancy resemble native speakers Readily understood by native speakers unaccustomed to non-native speakers
6	Advanced Low	Fairly organized paragraph- like speech with appropriate discourse markers (transitions, conclusion, etc.) Will not be as organized as level 7, but meaning is clear.	Can speak comfortably with concrete topics, and discuss a few topics abstractly Uses appropriate register according to prompt (formal or informal) Academic vocabulary often used appropriately	Grammar errors are infrequent and do not affect comprehension; no apparent sign of grammatical avoidance Able to speak in all major time frames, but lacks complete control of aspect Often able to successfully use compensation strategies to convey meaning Pausing resembles native patterns, rather than awkward hesitations Easily understood by native speakers unaccustomed to non-native speakers
5	Intermediate High	Simple paragraph length discourse with sustained, though possibly formulaic, discourse markers that help maintain some organization.	Able to comfortably handle all uncomplicated tasks relating to routine or daily events and personal interests and experiences Some hesitation may occur when dealing with more complicated tasks Uses a moderate amount of academic vocabulary appropriately	Uses a variety of time frames and structures; however, speaker may avoid more complex structures Error patterns may be evident, but errors do not distort meaning Frequent use of compensation strategies with consistent success Pronunciation problems occur, but meaning is still conveyed Exhibits break-down with more advanced tasks—i.e. failure to use circumlocution, significant hesitation, etc. Understood by native speakers unaccustomed to dealing with non-natives, but 1st language is evident
4	Intermediate Mid	Uses moderate-length sentences with simple transitions to connect ideas. Sentences may be strung together, but may not work together as cohesive paragraphs.	Able to handle a variety of uncomplicated tasks with concrete meaning Expresses meaning by creating and/or combining concrete and predictable elements of the language Uses sparse academic vocabulary appropriately	Strong command of basic structures; error patterns with complex grammar Frequent use of compensation strategies with varied success Pronunciation has significant errors that hinder comprehension of details, but not necessarily main idea Frequent pauses, reformulations and self-corrections Generally understood by sympathetic speakers accustomed to speaking with non- natives

3	Intermediate Low	Able to express personal meaning by using simple, but complete, sentences they know or hear from native speakers.	Able to successfully handle a limited number of uncomplicated tasks Concrete exchanges and predictable topics necessary for survival/(everyday life without unexpected complications) Uses highly varied general vocabulary	Errors are not uncommon and often obscure meaning Limited range of sentence structure Intonation, stress and word pronunciation are problematic and may obscure meaning Characterized by pauses, ineffective reformulations and self-corrections Generally be understood by speakers used to dealing with non-natives, but requires more effort
2	Novice High	Short and sometimes incomplete sentences.	Restricted to a few of the predictable topics necessary for survival (basic personal information, basic objects, preferences, and immediate needs) Relies heavily on learned phrases or recombination of phrases and what they hear from interlocutor	Attempt to create simple sentences, but errors predominate and distort meaning Avoids using complex/difficult words, phrases or sentences Speaker's 1st language strongly influences pronunciation, vocabulary and syntax Generally understood by sympathetic speakers used to non-natives with repetition and rephrasing
1	Novice Mid	Isolated words and memorized phrases.	Rely almost solely on formulaic/memorized language Two or three word answers in responding to questions Very limited context for vocabulary	Communicate minimally and with difficulty Frequent pausing, recycling their own or interlocutor's words Resort to repetition, words from their native language, or silence if task is too difficult Understood with great difficulty even by those used to dealing with non-natives
0	Novice Low and less (non-functional)	Isolated words.	No real functional ability Given enough time and familiar cues, may be able to exchange greetings, give their identity and name a number of familiar objects from their immediate environment	Cannot participate in true conversational exchange Length of speaking sample may be insufficient to assess accuracy May be unintelligible because of pronunciation Nearly incomprehensible even by those used to dealing with non-natives

Appendix D: Details of the Examinee Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu Examinees
35	16	2.19	2.17	-5.26	.43	.98	.0	.97	.0	1.05	.61	.65	18 311
18	8	2.25	2.22	-5.10	.60	.55	-.9	.56	-.9	1.49	.88	.74	29 412
19	8	2.38	2.37	-4.72	.59	.82	-.2	.81	-.2	1.18	.26	.60	30 412
39	16	2.44	2.43	-4.55	.42	2.40	3.1	2.42	3.1	-.56	.01	.66	32 412
39	16	2.44	2.43	-4.55	.42	.64	-1.1	.64	-1.1	1.40	.74	.66	38 411
22	8	2.75	2.75	-3.66	.60	.73	-.4	.73	-.3	1.25	.65	.60	36 412
44	16	2.75	2.75	-3.65	.42	.84	-.3	.84	-.3	1.17	.79	.66	2 111
22	8	2.75	2.76	-3.65	.60	.20	-2.2	.19	-2.3	1.77	.80	.60	19 312
23	8	2.88	2.87	-3.33	.60	1.36	.8	1.36	.8	.66	.80	.74	37 412
46	16	2.88	2.87	-3.31	.42	.57	-1.3	.57	-1.3	1.43	.86	.66	31 412
48	16	3.00	2.99	-2.95	.42	.55	-1.4	.55	-1.3	1.44	.80	.65	34 411
48	16	3.00	2.99	-2.95	.42	.74	-.6	.76	-.6	1.23	.70	.65	35 411
48	16	3.00	3.04	-2.83	.42	1.36	1.0	1.37	1.0	.65	.62	.65	10 211
25	8	3.13	3.12	-2.59	.60	.32	-1.6	.31	-1.6	1.65	.93	.73	20 312
25	8	3.13	3.12	-2.57	.60	.77	-.3	.77	-.3	1.25	.52	.60	5 121
25	8	3.13	3.16	-2.46	.60	.41	-1.4	.41	-1.4	1.62	.79	.60	12 221
51	16	3.19	3.18	-2.42	.43	.94	.0	.95	.0	1.06	.50	.65	40 421
51	16	3.19	3.18	-2.41	.43	.58	-1.3	.59	-1.2	1.42	.81	.65	17 312
51	16	3.19	3.22	-2.29	.43	.75	-.6	.77	-.6	1.24	.81	.65	9 211
53	16	3.31	3.31	-2.04	.43	.87	-.2	.84	-.3	1.17	.71	.65	3 122
56	16	3.50	3.49	-1.51	.43	.74	-.7	.75	-.6	1.26	.71	.64	33 412
29	8	3.63	3.61	-1.16	.60	.66	-.6	.67	-.6	1.34	.73	.72	44 422
59	16	3.69	3.68	-.96	.43	.72	-.7	.72	-.7	1.27	.69	.65	39 422
30	8	3.75	3.74	-.79	.61	.90	.0	.91	.0	1.10	.59	.73	22 321
60	16	3.75	3.74	-.77	.43	1.20	.6	1.20	.6	.82	.54	.65	24 322
60	16	3.75	3.78	-.66	.43	1.09	.3	1.10	.3	.91	.88	.65	11 222
31	8	3.88	3.85	-.43	.61	.77	-.3	.76	-.3	1.23	.60	.73	43 421
31	8	3.88	3.87	-.38	.61	.54	-.8	.53	-.9	1.44	.27	.59	23 321
63	16	3.94	3.92	-.22	.43	1.34	.9	1.36	1.0	.67	.69	.65	41 422
63	16	3.94	3.92	-.22	.43	.95	.0	.96	.0	1.04	.54	.65	47 421
32	8	4.00	3.99	-.01	.61	.42	-1.2	.42	-1.2	1.56	.66	.58	42 421
65	16	4.06	4.04	.15	.43	.71	-.7	.70	-.7	1.28	.68	.65	21 322
33	8	4.13	4.11	.36	.61	.43	-1.3	.42	-1.3	1.58	.79	.58	26 332
67	16	4.19	4.15	.51	.43	.76	-.6	.76	-.6	1.23	.81	.65	46 422
34	8	4.25	4.21	.67	.60	1.68	1.2	1.67	1.2	.35	.44	.73	7 131
68	16	4.25	4.21	.69	.43	2.01	2.3	2.00	2.3	-.02	.12	.64	50 431
68	16	4.25	4.25	.81	.43	.52	-1.5	.52	-1.5	1.48	.77	.64	14 232
69	16	4.31	4.28	.88	.42	.94	.0	.92	-.1	1.09	.86	.64	1 132
70	16	4.38	4.33	1.04	.42	.47	-1.8	.47	-1.8	1.55	.74	.65	48 421
35	8	4.38	4.34	1.07	.60	.43	-1.3	.42	-1.3	1.60	.61	.59	45 422
35	8	4.38	4.35	1.08	.60	.79	-.2	.81	-.2	1.20	.67	.59	6 131
69	15	4.60	4.57	1.71	.43	1.13	.4	1.17	.5	.83	.49	.66	8 132
37	8	4.63	4.57	1.72	.59	1.04	.2	1.02	.1	1.01	.70	.73	52 431
37	8	4.63	4.58	1.74	.59	1.19	.5	1.23	.6	.77	.65	.73	4 121
37	8	4.63	4.59	1.77	.59	1.69	1.2	1.70	1.3	.33	.05	.59	55 432
37	8	4.63	4.62	1.84	.59	.92	.0	.90	.0	1.11	.61	.73	13 221
37	8	4.63	4.62	1.84	.59	.68	-.5	.68	-.5	1.35	.95	.73	15 231
76	16	4.75	4.71	2.10	.42	1.18	.5	1.14	.5	.85	.41	.65	49 432
38	8	4.75	4.76	2.24	.59	1.09	.3	1.07	.3	.93	.39	.60	16 231
78	16	4.88	4.84	2.45	.42	.63	-1.0	.64	-1.0	1.35	.53	.66	54 431
39	8	4.88	4.85	2.47	.59	1.23	.5	1.21	.5	.80	.53	.60	53 431
40	8	5.00	4.96	2.79	.60	1.03	.2	1.06	.2	.94	.37	.74	27 332
81	16	5.06	5.03	2.97	.42	.89	-.1	.89	-.2	1.12	.65	.66	58 431
81	16	5.06	5.03	2.98	.42	1.27	.8	1.28	.8	.71	.76	.66	25 332
82	16	5.13	5.09	3.15	.42	1.57	1.5	1.55	1.4	.39	.68	.66	51 432
82	16	5.13	5.09	3.15	.42	1.57	1.5	1.62	1.6	.42	.49	.66	57 432
82	16	5.13	5.10	3.15	.42	.64	-1.0	.64	-1.0	1.37	.73	.66	28 331
47	8	5.88	5.87	5.29	.62	1.18	.5	1.20	.5	.75	.39	.73	56 432
47.8	12.4	3.87	3.86	-.41	.50	.92	-.2	.92	-.2		.63		Mean (Count: 58)
18.5	4.0	.88	.87	2.53	.09	.42	1.0	.42	1.0		.20		S.D. (Population)
18.7	4.0	.89	.88	2.55	.09	.42	1.1	.43	1.1		.21		S.D. (Sample)

Model, Populn: RMSE .51 Adj (True) S.D. 2.48 Separation 4.85 Strata 6.80 Reliability .96
Model, Sample: RMSE .51 Adj (True) S.D. 2.50 Separation 4.90 Strata 6.86 Reliability .96
Model, Fixed (all same) chi-square: 1595.5 d.f.: 57 significance (probability): .00
Model, Random (normal) chi-square: 55.4 d.f.: 56 significance (probability): .50

Appendix E: Details of the Rater Measurement Report

Total Score	Total Count	Obsvd Average	Fair(M) Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Exact Agree. Obs %	Exact Agree. Exp %	Nu	Rater
136	45	3.02	3.07	-2.32	.25	.90	-.4	.91	-.3	1.09	.84	.83	9.4	7.5	14	11
137	45	3.04	3.13	-2.13	.26	1.57	2.3	1.57	2.3	.44	.87	.84	27.3	25.1	2	13
145	45	3.22	3.36	-1.47	.25	1.36	1.6	1.37	1.6	.59	.75	.84	.0	.0	7	33
169	45	3.76	3.62	-.72	.25	.83	-.7	.83	-.7	1.18	.87	.84	40.0	45.5	3	22
159	45	3.53	3.66	-.62	.25	.47	-3.0	.46	-3.1	1.53	.89	.83	48.9	40.3	10	12
159	45	3.53	3.71	-.47	.25	.80	-.9	.80	-.9	1.22	.78	.83	57.8	41.0	11	32
171	45	3.80	3.71	-.45	.25	.76	-1.1	.77	-1.1	1.23	.82	.84	45.9	38.8	5	23
182	45	4.04	3.90	.11	.25	.89	-.4	.88	-.4	1.11	.90	.84	40.0	45.5	8	22
180	45	4.00	3.97	.36	.25	.93	-.2	.93	-.2	1.08	.81	.83	48.6	44.4	15	23
181	45	4.02	3.99	.42	.25	.75	-1.2	.75	-1.2	1.25	.85	.83	51.4	44.4	12	23
175	44	3.98	4.05	.59	.26	1.35	1.5	1.36	1.5	.66	.89	.84	27.3	25.1	6	13
192	45	4.27	4.10	.74	.25	.86	-.6	.87	-.5	1.14	.81	.83	.0	.0	9	21
182	45	4.04	4.13	.84	.25	1.02	.1	1.00	.0	.99	.89	.83	48.9	40.3	13	12
181	45	4.02	4.16	.92	.25	.68	-1.6	.68	-1.6	1.33	.90	.83	57.8	41.0	16	32
208	45	4.62	4.48	1.88	.25	.67	-1.7	.66	-1.7	1.34	.86	.84	32.1	34.9	1	23
213	45	4.73	4.64	2.32	.25	1.30	1.3	1.32	1.4	.69	.77	.84	9.4	7.5	4	11
173.1	44.9	3.85	3.86	.00	.25	.95	-.3	.95	-.3		.84				Mean (Count: 16)	
21.5	.2	.48	.42	1.25	.00	.29	1.4	.30	1.4		.05				S.D. (Population)	
22.2	.3	.49	.43	1.29	.00	.30	1.4	.30	1.5		.05				S.D. (Sample)	

Model, Populn: RMSE .25 Adj (True) S.D. 1.23 Separation 4.85 Strata 6.79 Reliability (not inter-rater) .96
Model, Sample: RMSE .25 Adj (True) S.D. 1.27 Separation 5.01 Strata 7.02 Reliability (not inter-rater) .96
Model, Fixed (all same) chi-square: 389.7 d.f.: 15 significance (probability): .00
Model, Random (normal) chi-square: 14.5 d.f.: 14 significance (probability): .41
Inter-Rater agreement opportunities: 429 Exact agreements: 178 = 41.5% Expected: 159.1 = 37.1%