2010-07-16

Examining Rater Bias in Elicited Imitation Scoring: Influence of Rater's L1 and L2 Background to the Ratings

Min Hye Son
Brigham Young University - Provo

Examining Rater Bias: An Evaluation of Possible Factors Influencing Elicited Imitation Ratings

Minhye Son

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

Master of Arts

Dan P. Dewey, Chair
Deryle Lonsdale
Ray Graham

Department of Linguistics and English Language

Brigham Young University

August 2010

ABSTRACT

Examining Rater Bias: An Evaluation of Possible Factors Influencing Elicited Imitation Ratings

Minhye Son

Department of Linguistics and English Language

Master of Arts

Elicited Imitation (EI), which is a way of assessing language learners' speaking, has been used for years. Furthermore, there have been many studies done showing rater bias (variance in test ratings associated with a specific rater and attributable to the attributes of a test taker) in language assessment. In this project, I evaluated possible rater bias, focusing mostly on bias attributable to raters' and test takers' language backgrounds, as seen in EI ratings. I reviewed literature on test rater bias, participated in a study of language background and rater bias, and produced recommendations for reducing bias in EI administration. Also, based on possible rater bias effects discussed in the literature I reviewed and on results of the research study I participated in, I created a registration tool to collect raters' background information that might be helpful in evaluating and reducing rater bias in future EI testing. My project also involved producing a co-authored research paper. In that paper we found no bias effect based on rater first or second language background.

Keywords: Elicited Imitation, Rater bias, Language Assessment, Rating

ACKNOWLEDGEMENTS

I would not have been able to finish this project without the help of Dr. Dewey. I really appreciate his endless support, encouragement, and patience from the beginning to the end. I am also thankful for my committee members, Dr. Lonsdale and Dr. Graham for their great examples. Jerry McGhee gave me much help in accomplishing this project too. I would like to thank him for his continual help even with his busy schedule. Finally, I am really grateful for my family and friends. I would not be here without their infinite love, support, and faith in me.

Table of Contents

List of Tables

Chapter 1
Overview

My first exposure to elicited imitation (EI) was as an EI grader/rater at Brigham Young University (BYU) of Hawaii. At that time I worked with another rater (an American) and, as I regularly discussed my ratings with him, I realized that there can be considerable variation between raters, depending on their background and experience. My grading experience got me very interested in this topic and the project I will report here.

In what follows, I will give a description of my M.A. project. First, I will provide some background on Elicited Imitation and will give a brief review of literature on the topic. Following that, I will give a brief description of my project, followed by detailed sections on each of the components of the project. After the project description, I will describe the process and progress of my project, which involved both creating the Registration Tool (a tool for providing background information useful in decreasing rater bias) and co-authoring a research paper addressing connections between rater language background (native and second language) and learner native language. Next, I will talk about some of the things I learned while working on this project. Finally, I will talk about connections between classes I took at BYU-Provo and my project in order to provide the reader with a picture of the expertise developed in connection with this project over the course of my M.A. degree experience. At the end, I will attach a copy of the research paper I co-authored with Dr. Dewey and others, tables summarizing the review of literature on rater bias, and a printed version of the Registration Tool I created.

Chapter 2
Background on EI

Elicited Imitation (EI) is a technique for language testing which has been used for years. Although there have been some questions regarding the validity of the technique, it has been accepted by many researchers as a tool useful for a variety of purposes. Many studies have also shown high correlations between EI and the Oral Proficiency Interview (OPI; see http://www.languagetesting.com/ for more information on this test, created by ACTFL, the American Council on the Teaching of Foreign Languages) and other measures of speaking proficiency. Therefore, it has been considered as an effective way of quickly and roughly assessing learners' speaking proficiency. The technique used in EI consists of reading an utterance to subjects, who are then requested to repeat it as exactly as possible. Responses are then recorded for later grading. The more proficient a speaker is, the longer and more complex the sentences which he or she can accurately repeat will be.

The PSST (Pedagogical Software and Speech Technology) research group at the BYU Department of Linguistics and English Language is working on exploring and expanding the use of speech technology in language learning. This research group evaluates existing speech technology, examines pedagogical needs, and designs and develops improved technological tools for language learning. One of the projects this research group is focusing on is developing EI as an oral language testing technique, which is inexpensive, efficient, and reliable.

About 700 second language (L2) learners have participated in EI testing performed by PSST researchers. There are 60 items per test with four different test forms.

In the PSST's EI, two separate human raters score each sentence spoken by the test takers. Raters listen to each item and, using a computer-based interface, determine which syllables in each sentence students repeat correctly. There are typically a variety of raters in terms of native and second language backgrounds, but for the research paper that was part of this project there were twenty raters consisting of half native speakers of English and half non-native speakers of English. Raters came from a variety of language backgrounds, but all were either native English speakers or highly proficient second language speakers of English.

The specific project that I worked on here was investigating connections between rater attributes and the ratings they assigned for the EI. I also looked at the characteristics of the test takers and correlations between these characteristics and the ratings the test takers received. Raters' different backgrounds might influence the results of their ratings, which is an example of 'rater bias' in the sense that it is used in this write-up (see section entitled Rater Bias under Review of Literature below). The raters used in past PSST research have different backgrounds: some are native speakers of English and others non-native; some are speakers of Romance languages and others speakers of Asian languages; some have more experience working with English language learners and others have little experience; some may be more sympathetic than others. Even among non-native speakers of English, their L1 background, number of years studying English, English proficiency, and ages are different. Given this variety, I decided to focus my project and the related products on rater agreement and factors contributing to ratings. I did an extensive review of literature and worked on a co-authored piece of original research to see the inter-rater reliability among raters with

different language background. The results of this project will inform the PSST and other groups using the EI to measure learners' language abilities whether they need to consider raters' language background when they hire raters.

The purpose of this project was to learn about rater bias and ways of improving reliability. More specifically, it was to evaluate bias in the EI used by the PSST and to help reduce bias to increase reliability of EI scoring results. For my project, these are the things I worked on: 1) conducting a review of literature on rater bias, 2) identifying factors that create bias; 3) focusing my attention on one key variable of concern to the PSST group, rater language background (whether raters with different language background produce different scoring results), 4) creating a tool to help increase reliability, reduce rater bias, and facilitate research, and 5) making recommendations to the PSST group for maximizing reliability of EI scoring.

Chapter 3
Literature Review

To give readers a general idea of previous research findings on connections between rater backgrounds and the ratings they assign to test takers, I present here a brief review of some of the literature focusing on this topic. Specifically, I highlight three areas that are commonly addressed in research on bias: language background, experience working with English (L2) language learners, and rater training. To help understand these three areas, I define the concept of bias in greater detail. Additional definitions and references can be found in the draft of the co-authored study found in the appendix.

Rater Bias

Overall, rater bias is defined as variance not as overall leniency or severity of ratings (some raters can just tend to be hard on test takers overall and others much softer in general), but more in terms of systematic variance that can be associated in some way with test taker attributes, such as language background, age, gender, educational level, etc. Rater bias has been approached in two main ways. The first way is a more general conceptual way and the second is more technical and involves finding patterns in rater performance using statistical techniques. (Caban, 2003; Chaulhoub-Deville & Wigglesworth, 2005; Wigglesworth, 1994). The first approach usually involves comparing ratings for different groups (e.g., male vs. female, one type of student vs. another, etc.) by the same rater(s) and determining if ratings for the groups compared are significantly different from each other. The second way involves the use of FACETS and other statistical procedures to find patterns in rater performance and then trying to find explanations for those patterns that go beyond test taker performance (Eckes, 2005, 2008; Weigle, 1998; Wigglesworth, 1993), or what Eckes (2005) calls "consistent deviations from what is expected on the basis of the [statistical] model." (p. 203). Most of the work in this project deals with the approach to bias analysis.

Language Background

The first factor we will consider is raters' language background (first language and second language). It is possible that rater bias exists among raters with different language backgrounds. More specifically, for English language tests such as our EI, native speakers of English and non-native speakers of English might rate learners

differently, and raters who speak the native language of the test takers might also be biased. Wigglesworth (1994) conducted a study to explore rater bias in rating an oral interaction test, connecting particular tasks in a test with particular raters. She found that rater nationality did relate to the way they scored particular tasks, but the effect size was so small she estimated it was not worth worrying about. As she concluded her study, one question was raised in her mind: "whether raters *from [particular] countries* would be biased toward the native speakers of that country due to their own (that is raters) familiarity with the interlanguage and pronunciation of the candidates." (p. 89, italics added). Familiarity with certain languages may help raters to understand the languages better and give better scores than other raters who are not familiar with the languages. In contrast, that familiarity might make raters be harsher or less tolerant of the mistakes that test takers from the same language background make. In Brown's research (1995) to find out whether different types of raters perceive the items in a test differently, non-native speakers of a language were found to be harsher on certain items such as pronunciation than native speakers because of their experience learning the second language. On the other hand, Du, Wright, and Brown (1996) found no significant rater bias against student ethnic groups in their study. Myford and her colleagues (1996) found that the number of languages spoken by raters correlated with reader severity when evaluating the possible influence of rater background.

Among PSST raters who are non-native speakers of English, number of years studying English and English proficiency are also different. One of the key questions being addressed by this project is whether these factors influence raters' scoring or not.

Experience Working with English Language Learners

A few studies have been done showing bias between raters who have experience teaching the target language of the examinees and those who do not have such experience. First, Galloway (1980) had thirty-three raters evaluate the oral communication of ten students who were learning Spanish. The raters were divided into four groups according to their Spanish teaching experience and their first language (native and non native speakers of Spanish). The results showed there were no significant differences among the groups on ratings of informational communication. However, comments made during the rating process showed how differently each group perceived students' mistakes. While raters with teaching experience were more critical of pronunciation and rate of speed, raters who were native speakers of Spanish with no teaching experience were more generous on these aspects.

Another study done by Hadden (1991) addressed teacher and non-teacher perceptions of second-language communication. Both ESL teachers and non-teachers who were native speakers of English completed a questionnaire after viewing videotapes recorded by native Chinese speakers in an ESL class. They were asked to indicate their perceptions of the speaker's communication on five different dimensions: 1) linguistic competence, 2) comprehensibility, 3) personality, 4) content of the presentation, and 5) manner of communication. The results indicated that perceptions of teachers and non-teachers did not differ greatly, except on discrete linguistic abilities such as pronunciation. Compared to the teachers, the non-teachers were more tolerant on students' linguistic performance.

Given apparent differences between those with and without language teaching experience, we controlled for this variable in the co-authored paper. The PSST Group might consider researching the effects of this variable in EI ratings, in particular if they have raters evaluate pronunciation or other discrete linguistic variables.

Rater Training

Rater training has often been assumed to increase inter-rater reliability: the consistency of the results among raters. However, it is impossible to fully eliminate rater variability even after training. Little research has been done to find out the effectiveness of rater training. Research done by Weigle (1998), and Elder, Barkhuizen, Knoch, and Randow (2007) on rater training effects indicated that no big differences were shown in inter-rater reliability after rater training, but rather that the training helped to increase intra-rater reliability (consistency by an individual rater). Although this may seem counter-intuitive, the studies by Weigle and Elder and her colleagues seem convincing. Further research in this area may be needed. In Wigglesworth's (1993) study, thirteen raters participated in a first rating session. Then eight of the raters were called again to participate in a second rating session after a two-part refresher rater training. In that training, raters first received individual feedback on their ratings. Then, in the second session, a group rating-training session was held. The results showed that bias from the second rating session was reduced compared to the first rating session. Wigglesworth noted that providing feedback on raters' individual performance served to reduce bias. Similar training could be conducted by the PSST Group. For the co-authored research paper, we controlled for rater training (all had the same amount of training).

Chapter 4
Research Paper

There are two major final tangible products from this project. The first is a research paper prepared for publication in conjunction with Dr. Dewey and Jerry McGhee. The paper includes a review of literature, a description of research methods and results, and a discussion and conclusion. I wrote the first draft of this paper and Dr. Dewey and Jerry McGhee revised and added to the paper to prepare it for submission for publication. The anticipated venue is *Language Assessment Quarterly*, but other venues might be *Language Testing, Language Learning,* and *Educational Measurement: Issues and Practice.* We are also submitting a proposal to present our findings at the Second Language Research Forum at the University of Maryland in October.

Background

Questions the PSST group has had in the process of hiring EI raters were, 'can we hire both native speakers and non native speakers of English?' and 'will their ratings be the same?' Members of the group also wondered whether the nationality and first language background of the raters were important considerations. I chose to collaborate on a research project to address these questions.

Description of the Study

In order to find out the answers to the questions mentioned above, 20 raters who were native and non-native speakers of Japanese, Korean, Chinese, Spanish, and Portuguese (2 native and 2 non-native for each language) were selected to rate the EI test. These raters were assigned to rate the same 500 sentences repeated by 50 students from

our university's English Language Center (ELC). These 50 test takers were native speakers of Japanese, Korean, Chinese, Spanish, and Portuguese (equally distributed), which are the 5 most commonly spoken native languages of ELC students. For a more detailed description of the study, please refer to the co-authored paper in the Appendix.

Results of the Study

The results of this research showed that there was no significant interaction between rater language and student language. This means that there was not a systematic relationship between raters' language background and test takers' native language. Based on this result, it seems that considering the language background of raters may not be necessary in hiring EI raters.

Chapter 5
Registration Tool

The second major final tangible product from this project is the 'Registration Tool'. I created this Web-based tool to collect information on raters' background. The information collected by this tool is based on my review of the literature and the results of the collaborative study. I included all the possible background variables of raters considered to be possible contributors to 'rater bias,' potentially affecting test ratings. The registration system collects the following information regarding the rater: age, gender, native language, additional languages spoken and level of proficiency in those languages, and time spent teaching English as a second language. The Registration Tool will be used by the PSST to register future raters. The data input by raters can be used in future studies similar to the co-authored study included here to analyze possible bias

based on rater background. At present, the Registration Tool will be used to collect information from raters during rater recruiting and rater training sessions. Please see the Appendix for the copy of the registration tool (screen shots).

Chapter 6
Recommendations to the PSST Group

Working on this project gave me a chance to experience many different things: observing what professors and staff in the PSST research group do, being able to apply what I learned through my M.A. classes, learning new things about research, and so on. Reflecting on the things I experienced, I have some suggestions for improving the quality of the PSST's EI rating.

I participated in the pre-training provided to raters before they had started rating. It was one of the essential parts of my project for increasing reliability among raters. All raters received about 30-60 minutes of training and were introduced to a website where they could find answers to the questions they might have when rating on their own. The website lists possible questions raters might have and answers to those questions with some examples. During the training, raters received a brief explanation on what the purpose of this project was and how they should rate sentences. Then, they practiced rating a few random example sentences. While they were practicing and referring to the website to get the answers to questions as needed, they often had questions on terminology used on the website, such as morphemes and phonemes. Because most of the raters were not familiar with these terms, they could not fully understand the explanations on the website. Some raters could not clearly understand the explanations on the website for other reasons. I had a strong feeling that they might later face similar situations again

while rating on their own if the trainer would not explicitly go over each questions and answers addressed on the website and explain what they mean with examples. This could significantly improve raters' ability to find answers to their questions and thereby be more consistent in their ratings. Without such training, whenever raters had questions, they would have needed to either contact the trainer or follow their own interpretation and judgments, which would affect inter-rater reliability of ratings.

Discussing each of the question-answer pairs from the rater website could help improve reliability, but raters are bound to still have difficulty understanding the answers when they work through things on their own after training. For this reason, it would be good to revise the questions to make them more readable and rater-friendly. The PSST rater trainers could make detailed notes about questions raters have as they try to use the website. The PSST Group could also have current raters or people who would be potential raters in the future (people with traits typical of PSST EI raters) look at the questions and identify anything they feel is unclear. The questions and answers could then be revised to make them more readable. The PSST Group could also follow up to watch what raters do after they read answers to make sure they do what they are expected to do after they read the answers.

Regarding the training session I observed, as I watched, I thought of a way of providing more effective training. Here is a suggested outline for training. The training would be held in a lab with every rater having his/her own computer. The trainer would do the following:

1. Explain the background of this project and the process of the rating.

2. Show and explain the examples of screens they will see for rating.
   (Use Power Point, so everyone can be on the same track.)
3. Have the raters practice rating by themselves. (ask the raters to write down any questions.)
4. Talk about questions raters have together.
5. Explain to the raters how they can find answers to many of their questions from the website section 'Grading FAQ-PSST'.
6. Go through the questions listed on the website together. Prepare in advance a few example sentences with full recordings of the sentences for each question. Have the raters listen to the recording, identify the problems in the sentences (aiming for those written on the website), and find out how they need to rate based on what they read on the website.
7. Give raters time to practice grading items while referring to the FAQ section of the website.
8. Come back as a group to discuss questions raters had while working on their own.
9. Give several more random example sentences to practice to the group. Working on this as a group, they might have additional questions on how to rate which may not be addressed on the web site. Work toward consensus as a group, making sure raters follow the PSST guidelines.

Providing the training in this way will help the raters to minimize misunderstandings, confusion, or questions that might arise as they work on rating by themselves.

Another suggestion I would like to make is to conduct further research to find out other possible rater bias which might affect results of ratings such as gender and age of raters. This study shows no connections between rater language backgrounds and the ratings they give. Based on the literature review, there are other possible factors that might have an effect. Therefore, connections between ratings and rater backgrounds (age, gender, training, experience, etc.) should be researched. In order to do this, the PSST group needs to measure and/or control for these variables. Then, by having the raters rate the same data which was used for this study and analyzing their ratings, they will be able to find out how these variables influence ratings.

Chapter 7
My Project Efforts

In this part, I would like to describe the work that was involved in this project. At the end, I include a table summarizing the activities I was engaged in, hours I spent on each activity, and results/accomplishments of the activities.

Attending PSST Weekly Meeting

As I started working on this project with Dr. Dewey, I became a member of the Pedagogical Software and Speech Technology (PSST) Research Group. The PSST Research Group holds weekly one-hour meetings. I have been attending the weekly meetings since the summer of 2009. In the meetings, we first share what each person is working on, and then provide updates on our projects. Sometimes, questions are brought up from members regarding their specific projects. Then, we discuss these questions together to find good solutions for each other. There were about two or three times when

a few members gave presentations on their projects as practice for their presentations in upcoming conferences. After the presentations, group members provided feedback to enhance the quality of the presentations. I thought it was a great opportunity to help each other. Another activity we engage in during each meeting is the discussion of a research paper, which is related to our projects. One member of the group sends an article before the meeting, members read the articles, and we all discuss the paper together. This helps me to extend my knowledge on the projects we are working on beyond just my own project. Attending this meeting also helps me to know what professors do, besides teaching, to promote progress in their fields. Before, I was glad thinking that, once I graduate and get a job, I would not need to study any more. However, I realize that I was wrong. I need to continually work on learning and expanding my knowledge on this area. This will not only help me to progress personally, but it will also benefit the people with whom I am involved, such as my students and colleagues.

Finding Research Papers on Rater Bias

One of the most important steps working on my project was finding articles related to my project and writing a literature review in order to inform the PSST Research Group. At first, it was kind of hard to find articles on rater bias (in particular related to language background), so I worked with Dr. Dewey to determine people in the testing field that I could contact to get some information regarding possible references I was not finding. I was able to get some helpful resources from some of them. I also continually searched the internet, journals, and databases to get as much information and as many research papers as possible. Through this additional effort, I was able to find more good

resources and become familiar with a variety of databases and search engines. Reading the research papers provided additional references and names of researchers interested in the topic of bias. As I found more articles, read them, and got more information on rater bias, I became more fascinated with the topic of bias. After reading the research papers, I made a chart and summarized each paper in that chart, including the name of the research(s), purpose(s) of the research, methods, results, and conclusions. This helped me later when I worked on writing up the literature review for this write-up and for the co-authored paper.

Hiring, Training, and Supervising Raters

I spent a lot of time working with raters. For the joint research paper, Dr. Dewey and I decided to hire two native speakers and two non-native speakers for each of the languages of interest, which are Japanese, Korean, Mandarin, Portuguese, and Spanish, (total twenty raters). I first wrote a job description for the advertisement. I was the contact person for the entire hiring process. In this process, collected extensive information about each person (age, language background, gender, experience living abroad, experience teaching language, etc.) because I wanted to have consistency among raters. Because our ratings occurred right before winter vacation, it was difficult to find some raters for certain languages. There were about 15 raters who started at the beginning of the winter vacation. Before they started rating, there were training sessions. I worked with the PSST members in charge of the EI system to train all of the raters. Because I did not initially know how to train raters, I attended initial training sessions to learn how myself. There were raters who joined later, so I met with them later for

individual training. There was much confusion over winter break because of a technical problem we faced. So, since raters had already started rating, I had to be continually in contact with them through e-mails and phone calls to make sure everyone was on the right track. There were a few raters who had to quit, so I had to find more raters who could replace them and then had to train these new raters. I worked with Lorianne Spear, Secretary in the Linguistics Department, to complete the necessary paperwork to hire all of the raters. From this process, I learned how to be an effective and well organized supervisor, how to take care of the logistical issues related to hiring, and how to work through challenges associated with carrying out a program that involves both technical and personnel challenges.

Meetings with Dr. Dewey and PSST Staff

I have been meeting with Dr. Dewey and Jerry McGhee, the key member of the PSST in charge of adding raters to the EI grading system and setting up specific grading profiles for projects such as ours. I have been in continual contact with Dr. Dewey and Jerry through e-mail to get the grading system set up, add and train raters, and troubleshoot as problems came up. These two people are the ones who have been helping me the most to be able to continue working on my project. I met with Dr. Dewey at least once or twice a week for about 30 – 60 minutes each time. We met during the whole time I worked on my project. At the beginning of my project, I had to meet with Jerry a few times a week and continually contact him through e-mails to get his insights and assistance. When we faced technical problems, I had to contact him several times a day.

Jerry assisted me later in the project by helping me retrieve and organize the rating results.

Analyzing the Data

Dr. Dewey and I met with a statistician to analyze the data we got from our raters. I learned that it is important to have all your data in proper order for analysis. We found some initial problems with the data and had to go back and add and re-code data before we were able to conduct our final analysis. It was a great opportunity to see how statistical procedures I learned from my testing class can be applied in language assessment research. Through this experience, I developed a better understanding of statistical concepts such as ANOVA, variable types (random, fixed, nominal, interval, etc.), correlation, and mixed linear modeling.

Summary of Time Spent

Table 1 shows the approximate amount of time spent in activities related to this project. These are estimates based on reflection after completion of the project.

Table 1.

Time Spent on Project-Related Activities

Activity	Hours	Results/Accomplishments
Attending PSST Meetings	25 – 30 hours	Getting professionally involved in a research group
Doing Literature Review	25-30	Acquiring and developing an understanding of

	hours	research on rater bias
Hiring, Training, and Supervising Raters	25 hours	Helping raters to successfully finish rating
Meeting with Dr. Dewey and PSST Staff	25hours	Getting advice and assistance and sharing updates on research progress (review of literature and research paper)
Analyzing the Data	4 hours	Meeting with statistician and calculating the results of our research
Working on Papers	50 hours	Writing 'Write – Up' and 'Co-Authored Paper'

Chapter 8
Connections Between Coursework and My Project

Classes I took here at BYU and skills I gained from the classes helped me significantly while I was working on this project. In this section, I will discuss connections between my project and classes I took (classes taken as part of M.A. core curriculum and additional classes taken to support completion of the M.A. project).

The first Linguistics class I took at BYU was 'Introduction to Research in TESOL.' In this class, I learned how to analyze and interpret published research for language teachers and researchers. As a major assignment for the class, I had to write a review of literature on a topic I was interested in at that time. So, using the skills I learned from the class, I found articles on a specific topic, analyzed them, and then, wrote a

literature review. In the process, I was exposed to many research papers and it helped me to understand how research papers are formatted, how research on language learning is designed, and how specific parts of research papers are written. This greatly helped me when I co-authored the research paper for my project.

As I completed the TESOL Certificate program, I took four credits of 'TESOL Practicum.' The purpose of this class is to help students with actual fieldwork experiences in TESOL settings. I did an internship at the ELC, teaching a grammar class and, since then, I have been continually working at the ELC, teaching grammar, writing, and oral communication classes. There are about 180 students from 30 different countries at the ELC. This helped me to see how students from different countries learn and speak English, and to perceive what they learn differently. More important for this project, I became more aware that there are patterns in L2 learner language that are often common to learners with the same L1 backgrounds. I also learned more about language assessment. All the teachers at the ELC are required to assess students' language abilities during achievement tests given at the end of each semester. Before we start rating, we always receive training (even if we have received the training before) on how to rate in order to maximize reliability). Rating the assessment, I often thought about how the results of the ratings would be different depending on the different backgrounds of the teachers. The Elicited Imitation test, used in the research conducted as part of the co-authored paper that was part of this M.A. project, is a part of their speaking assessments given at the end of each semester at the ELC (though ELC teachers do not rate this test). One thing that attracted me in this project was my teaching and rating experiences at the ELC.

Taking the 'Technology in Language Teaching' class helped me to broaden my views on the use of technologies and to develop skills helpful in producing the Registration Tool (a Web-based survey), organizing and analyzing the data (using Excel and working with statistics on a computer, etc.), and training the raters. The elicited imitation was created to assess language learners' speaking using technology in an inexpensive, efficient, and reliable way. While taking the Technology in Language Teaching class and working on my project at the same time, I was able to more closely see how technology can be used in language learning, in particular in assessment. In short, the class made me much more comfortable with the technology tools used by language teachers and researchers, facilitating completion of my project.

I also took the 'Language Testing' class, learning various methods for assessing language skills, and learning about construction, analysis, use, and interpretation of language tests. I learned of the importance of reliability, a concept that was at the center of this project. Since my project focused specifically on the inter-rater reliability of the Elicited Imitation, I could more easily relate and apply what I was learning in class to my project. I was also introduced to statistical procedures necessary for evaluation of language tests (e.g., Spearman rho). Although I had once briefly learned statistics when I was attending BYU-Hawaii, statistics seemed very hard for me and I did not fully understand why I had to learn statistics when I want to be a language teacher. However, I later realized through this class and my project how to apply some of these concepts. In order to analyze the ratings of the raters who participated in my project, Dr. Dewey and I met with a statistician several times. This provided some hands-on experience with the statistical procedures studied in class. If I have opportunities to use statistical procedures

(in particular in test analysis) in the future, I am certain that this class and experience will be a great foundation for additional work in statistics.

As many of us in the TESOL M.A. program gain teaching experience, there might be some who end up being supervisors or administrators of language institutions. In preparation for such experience, there is a 'TESOL Supervision Administration Internship' class which provides us with actual fieldwork in TESOL settings involving supervision, in service training, and program administration. Taking this class, I had a chance to be a mentor for two students in the TESOL certificate program doing student teaching at the ELC and another community ESL program. This class and experience helped me to learn the qualities and skills supervisors and administrators need to have. As a big part of my M.A. project, I had to hire and supervise 20 raters. It was a more complicated process than I had thought it would be. Although I struggled to supervise and support these raters, in the process I learned what qualifications and skills I need to have and increase to be a better supervisor or administrator later on. I truly believe that a well prepared supervisor or administrator can make a big difference in his or her employees' attitude and passion towards their work, and the atmosphere of the work place.

Throughout the last two years, I have learned so many things I need to know to be a good language teacher such as language testing, technology, statistics, and administrative skills. Not only did I learn those things, but also I was actually able to apply what I learned through working on my project.

Chapter 9
Conclusion

I have been involved with Elicited Imitation since 2005, starting as a rater and now as a researcher looking at inter-rater reliability. When I worked as an EI rater at BYU-Hawaii, I often saw differences between my ratings and my co-worker's ratings, but I did not imagine that I would later work on this project to see how the differences I saw would affect the overall EI results. Working on this project has been a great opportunity for me in several ways. First, I was able to find out that the language background of the EI raters does not seem to affect the results of their ratings. Through this finding, I realized that the differences I had with my co-worker at BYU-Hawaii were not likely from our different language backgrounds. It was also interesting to find out the effects of test taker L1, gender, and level and to consider further research on these and other areas that might contribute to differences in ratings. Second, it helped me to broaden my experience and my perspective in my major. When I first decided to study TESOL, my main focus was on getting a better job after graduation. However, the more I got involved in this project and the PSST research group, and worked on my project, the more my passion toward my field increased and my desire for personal professional development increased as well. I have come to think more about what I can do after I graduate to continually be involved in this field. This leads to what I learned next. Researching and working on my project, I contacted and worked with many different people, including members of the PSST research group, professors, statisticians, and so on. I also spent a lot of time finding research papers on EI and rater bias. These

experiences helped me to see how many people are needed to complete a research project and how each one of their roles contributes to fulfillment of project goals. Lastly, I was able to see how my classes helped me to successfully accomplish my project. Working on this project, I combined pieces that I learned from my classes with new skills that I learned through mentors and PSST group members and applied these pieces to complete my project. Once again, I am truly grateful for all the things I learned through my project and all of the people who helped me out with their love and patience. Remembering and applying things I learned through this experience, I hope I can continue to develop as a TESOL professional.

References

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

Du, Y., & Others, A. (1996). *Differential facet functioning detection in direct writing assessment*

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

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

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.

Galloway, V. B. (1980). Perceptions of communicative efforts of American students of spanish. *Modern Language Journal, 64*(4), 428-33.

Hadden, B. L. (1991). Teacher and non teacher perceptions of second language communication. *Language Learning, 41(1),* 1-24

Myford, C. M., Marr, D. B., Linacre, J. M. (1996). Reader calibration and its potential role in equating for the test of written English: *Research Report 52*. Princeton, NJ: Educational Testing Service.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

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.

Appendix A

Summary of Studies on Rater Bias

Author(s) and Year	Article Title	Purpose	Candidates	Raters	Results	Conclusion
Anne Brown (1995)	The effect of rater variables in the development of an occupation-specific language performance test	Effect background has on assessments made on both linguistic and 'real-world' criteria	a. experience as a tour guide b. no experience	a. experienced tour guides b. experienced teachers of Japanese c. non-native speakers	a. NNS just follow guideline/ NS follows their intuition **b. Harshness** NS>NNS Indu>Teaching (but, minimal difference) c. NS and Indu background raters show more diverse harshness **d. Items:** Teachers being harsh on linguistic items VS. Indu raters being harsh on pronunciation NNS being harsher	The way they perceive items and apply the scale is different

Author(s) and Year	Article Title	Purpose	Candidates	Raters	Results	Conclusion
					on politeness and pronunciation Task fulfillment Indu raters and NNS being more lenient than teachers and NS e. **rating scale:** Teachers being reluctant on giving extreme scores	
Edward Schaefer (2008)	Rater bias patterns in an EFL writing assessment	Rater bias patterns of NES in rating EFL essays	Japanese EFL essays	NES raters (language teachers) living in Japan (1month to 4 years – 13.7 months)	(no significant effects for order, gender, and nationality) a. in**experienced** raters with minimal training used the scale to a satisfactory standard b. **by categories** (ex: content/organization /fluency pg. 475) most part within acceptable bounds, with few exception pg.482)	

Author(s) and Year	Article Title	Purpose	Candidates	Raters	Results	Conclusion
					c. **rater-category interaction** (being severe towards one, then lenient towards the other) d. **rater-writer interaction** (being more severe or lenient towards higher ability writers , because of high expectation?? – few exception)	
Gillian Wigglesworth (1993)	Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction	To investigate inter- and intra-rater reliability in the assessment of two alternate versions of an oral interaction test **"Rater**	94 ESL learners	13- rated tape 9 - rated 36 items each (half direct version tape/ half semi direct version tape) 4- all 166 tapes After feedback	Depending on the raters a. intra-rater reliabilities differ b. Harshness on specific items differ	Raters incorporate performance feedback – bias is reduced . . . Further research: whether the

Author(s) and Year	Article Title	Purpose	Candidates	Raters	Results	Conclusion
		Training"		and training sessions, all tapes were rated by two different raters		feedback is maintained over time
Tom Lumley and T.F. McNamara (1995)	Rater characteristics and rater bias: Implications for training	To investigate the use of these analytical techniques in rater training for the speaking subset To establish 1) consistency of rater characteristics over different occasions 2) rater bias in relation to occasion of rating	Occupational English Test to members in 11 different Health profession	**1.** Rater Training session (18 months later) **2.** Rater Training session (2 months later) **3.** Test Administration 13 Raters (Group A)- 1&2 4 Raters	→ Harshness Significant variations in rater harshness were shown. →Harshness Significant variations in rater harshness were shown. = Candidates are being measured consistently by the two groups of raters on the two occasions.	Reasonable consistency for all raters within these rating periods, with the significant variation coming over much longer periods. (pg. 68-69) -Substantial variation in rater harshness, which training has by no means eliminated, nor even reduced to a level which permit reporting **Results of training may not endure for long

Author(s) and Year	Article Title	Purpose	Candidates	Raters	Results	Conclusion
				(Group B)- 1,2,&3		after a training session ** Every administration, new calibrations of rater characteristics are required
Sara Cushing Weigle (1998)	Using FACETS t model rater training effects	To explore differences in rater severity consistency amo inexperienced ar experienced rate both before and after rater traini	60 essays from UCLA's English as a second language placement examination	16 raters 8 inexperienced - New (0-10 years of teaching experience) Female Like NS 8 experienced - Old (2-10 years of teaching experience)	**1.Rater Severity** New > Old : Both raters differ in their sev **2.Rater Consistency** Old> New : Untrained raters tend to be less reliable than trained raters **3.Post** (after training) a) making finer distinctions ar ability level b) raters are clustered together c) not much	Before training: Inexperienced rate being more severe less consistent After training: Differences were l pronounced, but s differences exist - training had the of reducing the extremism of the New raters

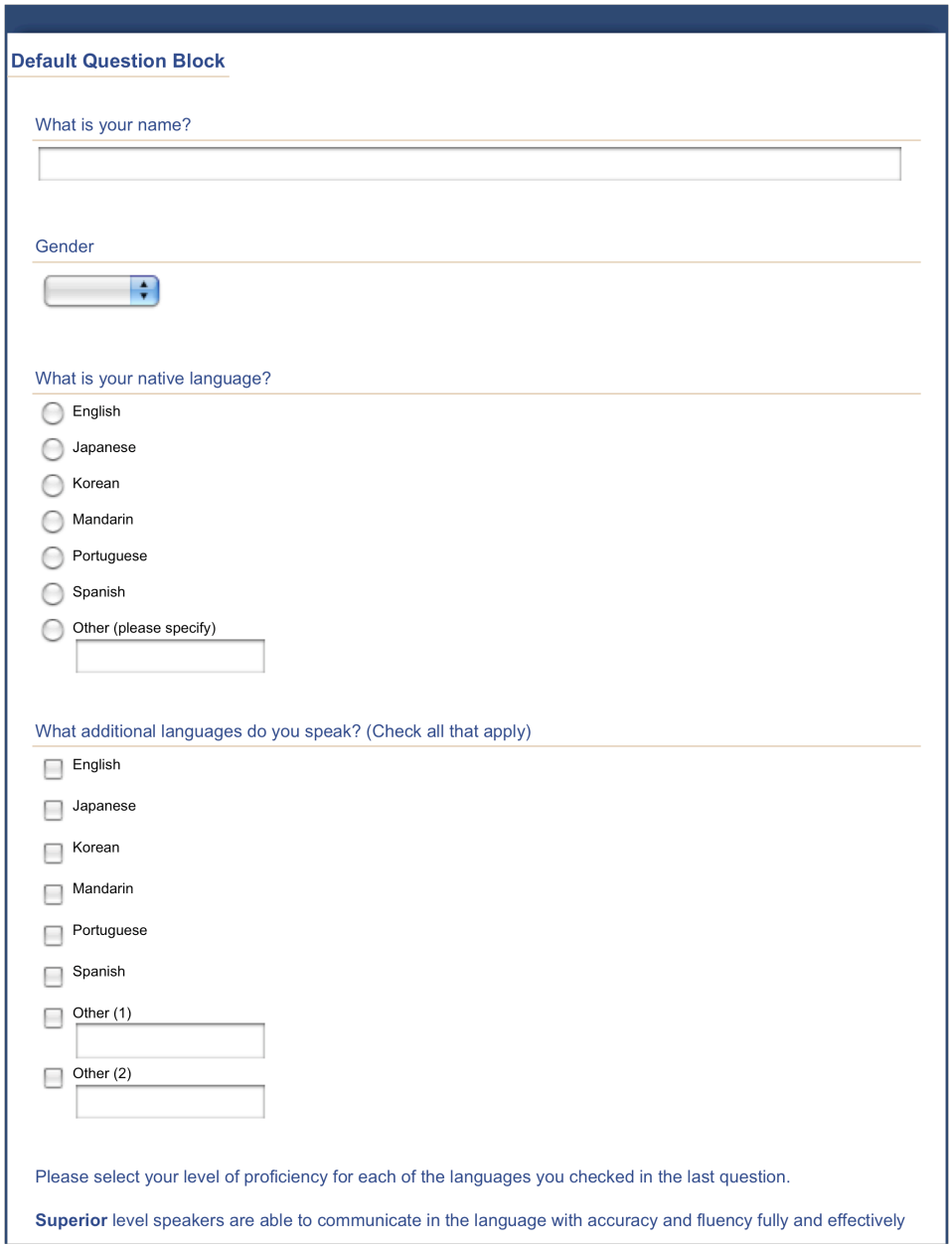
Author(s) and Year	Article Title	Purpose	Candidates	Raters	Results	Conclusion
					variation d) scales appear to be close together e) no clear distinction between new and old **4.Rater Consistency**	
Gillian Wigglesworth (1994)	Patterns of rate behavior in the assessment of an oral interaction tes	To determine whether any bias evident in the way a grou raters rated two different version an oral interactic test: Direct VS. Semi D Especially on Raters VS. Item : particular tasks were scored in a consistently biased way	83 candidates	13 raters on both oral interaction tests: live VS. tape version *FACETS:* Candidates Rater Test type Test order Item	1.Depending on the different versions of the test : Live VS. Taped 2. Specific assessment criteria :Grammar, Fluency, vocabulary - Raters lack consistency on certain items or tasks	One group rated tape-version more harshly while the group rated live version more harshly *Further Research* a. NS VS. NNS b. Particular language or cultural backgroun c. Gender
Catherine	Evaluating Rat	To investigate ra		8 raters	1.Online	-Slightly higher lev

Author(s) and Year	Article Title	Purpose	Candidates	Raters	Results	Conclusion
Elder Gary Barkhuizen Ute Knoch Janet Von Randow (2007)	Reponses to an online training program for L2 writing assessr	reactions to the online training Online Training f a) overcomi accessing training b) enhancing reliability c) impact on inter- and intra-rate consisten		a) experienced teachers b) not all NS, but with high levels of English proficiency c) experience of assessment but, a)different background b) different educati and assessment communities	Program: Opinions differ 2.Higher levels of inter-rater agreement followed by training : evenly distributed across all raters in group 3. Internally more consistent followed by training : change in intra- rater consistency is minimal	of overall inter-rat agreement -Reduced levels of inconsistency and bias in some instances Between online training and face-to face trainir which one is bette Not addressed in t study
Kimi Kondo-Brown (2002)	A FACETS anal of rater bias in measuring Japanese secon language writing performance	1) depending on candidates and criteria in assess writing, raters ar biased? :Rater-Candidate Interaction 2) exploring rati scale for norm	234 essay samples (78 samples on each of three prompts) from 234 students - Japanese language Proficiency	3 trained teacher rater	1. Severity differences among raters: significant variation in harshness did exist among the raters all raters were self consistent	Highly correlated scores and self-consistent But, Differences in ove severity (dependi on certain candidates

Author(s) and Year	Article Title	Purpose	Candidates	Raters	Results	Conclusion
		referenced decis	placement tests a) many female b) freshman wit different majors c) NS of English d) studied Japanese at leas one year e) beginner -lev		2. The difficulty and consistency for the five catego : the overall bias patterns identified between categories were consistent. 3. Rater bias in candidate abili a) within particular ability group, harshly or leniently bia interactions by a particular ra b) percentage of bias interaction was much higher for the candidates of extreme high or low ability 4. Rater bias in category difficulty : raters were consistent in the identified patterns of bias across all candidates	criteria) 1.The differences i severity between trained teacher ra were small but, significa 2. The difference i difficulty between most harshly scor category (organization) and the most lenie scored category (mechanics) was s but again significa
Thomas	Examining Rat	1. Rater VS. Face	Writing	Writing		1. Differ strongly i

Author(s) and Year	Article Title	Purpose	Candidates	Raters	Results	Conclusion
Eckes (2005)	Effect in TestDa Writing and Speaking Performance Assessment: A Many Facet R Analysis	- Examinees - Criteria (writi - Tasks (speaki 2. Rater VS. Genc	1359 participan (747 females, 612 males) Speaking 1348 participan (741 females, 607 males) from different countries	29 raters (23 women and 6 men) Speaking 31 raters (26 women, 5 men)		severity with whic they rated examin 2. Fairly consisten their overall ratin 3. Less consistent relation to rating criteria (or speaking tasks than in relation to examinees 4. No gender bias

Appendix B
Registration Tool



Default Question Block

What is your name?

Gender

What is your native language?

- English
- Japanese
- Korean
- Mandarin
- Portuguese
- Spanish
- Other (please specify)

What additional languages do you speak? (Check all that apply)

- English
- Japanese
- Korean
- Mandarin
- Portuguese
- Spanish
- Other (1)
- Other (2)

Please select your level of proficiency for each of the languages you checked in the last question.

Superior level speakers are able to communicate in the language with accuracy and fluency fully and effectively

participating in conversations on a variety of topics.

Advanced level speakers are able to handle a number of communicative tasks, but patterns of error appear.

Intermediate level speakers are able to converse with ease and confidence, but are unable to sustain performance at that level over a variety of topics.

Novice level speakers are able to communicate minimally and with difficulty relying heavily on learned phrases or recombinations of these and what they hear from their interlocutor.

	Novice	Intermediate	Advanced	Superior
» English	○	○	○	○
» Japanese	○	○	○	○
» Korean	○	○	○	○
» Mandarin	○	○	○	○
» Portuguese	○	○	○	○
» Spanish	○	○	○	○
» Other (1)	○	○	○	○
» Other (2)	○	○	○	○

What school year are you in?

- ○ Undergraduate Freshman
- ○ Undergraduate Sophomore
- ○ Undergraduate Junior
- ○ Undergraduate Senior
- ○ Graduate 1st year
- ○ Graduate 2nd year

Do you have English teaching experience?

- ○ Yes
- ○ No

Please describe your teaching experience, including the amount of time you taught and the nature of your experience.

Have you done Elicited Imitation rating before this?


Yes
No
For those of you who have done Elicited Imitation rating previously, did you receive training on how to rate? If so, please describe the training (when, where, and what the training was like)

Appendix C

Co-Authored Research Paper

Examining Rater Bias in Elicited Imitation Scoring: Influence of L1 and L2 Background

Minhye Son, Dan P. Dewey, Jeremiah McGhee, Dennis Eggett

Brigham Young University

Abstract

In this study we evaluated the possible influence of rater native and second language on their evaluations of performance on items from an elicited imitation (EI) test. Twenty raters, half native speakers of English and half highly-proficiency second language speakers, rated samples (10 EI items each) of speech produced by L2 speakers of English who were native speakers of Chinese, Japanese, Korean, Spanish, and Portuguese. Raters were also either native or second-language speakers of these same languages. No relationship was found between rater native or second-language and test taker native language, indicating that rater language background does not appear to bias their evaluations of test takers in EI.

Examining Rater Bias in Elicited Imitation Scoring: L1 and L2 Background

Studies of test rater behavior have regularly discovered unwanted variability due to the characteristics of the rater. Among the rater characteristics explored are experience teaching the language being tested (Caban, 2005; Eckes, 2008; Hadden, 1991; Galloway, 1980; Tanaka, Hajikano, and Tsubone, 1998), amount of training in rating (Barret, 2001; Weigle, 1998; Wigglesworth, 1993), amount of experience rating (Myford, 1996), test taker proficiency level (Schaefer, 2008), gender (Eckes, 2005; Schaefer, 2008), age (Eckes, 2008), rater native language (Caban 2003; Galloway, 1980), and second languages(s) spoken by the rater (Myford, 1996).

Investigating rater effects on second language writing performance assessment, Eckes (2008) found that age of rater and number of years of teaching the second language (German) were positively correlated with measures of severity. Specifically, the older a teacher, the more severe s/he was found to be in terms of evaluating the structure of the essays; the more years of experience teaching German a rater had, the more severe they were overall in rating the essays. Eckes also found that learners reported varying profiles in terms of what factors they weighed most heavily in their evaluations. These profiles were partially dependent on age, number of foreign languages spoken, and experience scoring writing.

Tanaka, Hajikano, and Tsubone (1998) also explored teaching as a variable influencing the rating of writing samples of learners of Japanese as a second language and, while teachers and non-teachers were found to give similar overall ratings to writing

samples, they disagreed on the importance of some criteria (importance of accuracy, ease of reading, use of certain sentence patterns and characters, etc.). Teachers were found to value accuracy more than non-teachers, a finding similar to Eckes (2008).

Exploring the role of teaching experience and native language in rating spoken language samples, Galloway (1980) had raters evaluate video-taped oral responses to speaking prompts for students who were learning Spanish. The raters were divided into four groups according to their Spanish teaching experience and their first language (native and non native speakers of Spanish). Results showed there were no significant differences among the groups on ratings of informational communication. However, comments made during the rating process show how differently each group perceived students' mistakes. While raters with teaching experience were more critical of pronunciation and rate of speed, raters who were native speakers of Spanish with no teaching experience were more generous on these aspects.

Another similar study by Hadden (1991) focused on teacher and non-teacher perceptions of second-language communication. Both ESL teachers and non-teachers who were native speakers of English completed a questionnaire after viewing videotape recordings of native Chinese speakers in an ESL classroom. They were asked to indicate their perceptions of the speaker's communication on five different dimensions: 1) linguistic performance, 2) comprehensibility, 3) personality, 4) content of the presentation, and 5) manner of communication. The results indicated that perceptions of teachers and non-teachers did not differ greatly, except on one dimension: compared to the teacher, the non-teachers were more tolerant of problems with students' linguistic performance.

In a small-scale study (rating the speech of only four individuals), Caban (2003) explored bias due to native language background in the rating of ESL oral interviews using a FACETS analysis. She found no significant differences between native and non-native raters and raters with and without teaching experience. Brown (1995) also evaluated whether different types of raters perceive the items in a test differently and found that non-native speakers of a language were harsher on certain items such as pronunciation than native speakers because of their experience learning the second language.

Evaluating the possible influence of rater background on scoring the Test of Written English (TWE), Myford and her colleagues (1996) found that number of foreign/additional languages spoken by the rater correlated significantly with reader severity: the more additional languages a rater spoke, the more severe they were in their ratings. Furthermore, those who had participated in more TWE rating sessions tended to me more sever in their ratings. Myford also found that "Stability of measures of severity was significantly related to number of years experience as a TWE reader . . . and to the number of languages known (i.e., the more languages the reader knew, the less stable the reader's two measures of severity)." (p. 40)

Examining patterns in rater evaluations of a writing and speaking performance assessment for speakers of German as a second language, Eckes (2005) found no trends in rater performance related to gender. However, he did find other patterns in rater performance. First, raters varied in their overall ratings, with some raters consistently scoring more severely or leniently than others. Second, he found that raters tended to weigh criteria differently in their ratings, even though overall ratings may have been

similar. These patterns were not associated with gender or other demographic variables. Eckes' finding of no gender bias is corroborated by Schaefer (2008), who also found no significant effects for gender in the rating of ESL essays.

One other background variable found to contribute to rater variability is amount of experience rating. Myford and her colleagues (1996) found that the more TWE ratings a person had performed, the more severe they were in their ratings. Furthermore, the more years a person had acted as a rater, the more stable they were in their ratings, regardless of their overall tendency in terms of severity. We are unaware of similar research regarding the rating of spoken language.

One other variable found to contribute to rater bias is test taker language ability. Schaefer (2008) found that some raters were stricter with second language writers with higher levels of writing proficiency than with lower-level writers, and other raters tended to be less severe with these same writers. Lower-ability writers were often rated more leniently on criteria by certain raters than higher-level writers. Whether similar patterns are seen in rating speech samples has yet to be determined.

Rater training has often been assumed to increase inter-rater reliability. However, it is impossible to fully eliminate rater variability even after training (Barret, 2001, Eckes, 2005; Lumley & McNamara, 1995). Research done by Weigle (1998), and Elder, Barkhuizen, Knoch, and Randow (2007) on rater training effects indicated that no significant differences existed in inter-rater reliability after rater training, but rather that the training helped to increase intra-rater reliability. In Wigglesworth's (1993) study, 13 raters participated in a first rating session. Then, 8 of the raters were called again to participate in a second rating session after a two-part refresher rater training. In that

training, raters first received individual feedback on their ratings. Then, in the second session, a group rating training session was held. The results showed that bias from the second rating session was reduced compared to the first rating session. Wigglesworth noted that providing feedback on raters' individual performance using bias analysis served to reduce bias significantly. In a subsequent study, Wigglesworth stated that the research to date suggests, "rater training sessions can address some of the concerns of rater variability but not others, and that controlling for rater background may reduce variability." (p. 78).

In an attempt to identify possible bias based on rater's cultural backgrounds, Chalhoub-Deville & Wigglesworth (2005) compared raters from Australia, Canada, the U.K., and the U.S. evaluating performance by ESL learners on an oral interaction test. Chalhoub-Deville & Wigglesworth concluded, "Analyses performed show a statistical difference among the groups in their ratings of test takers' oral performances for each of the three tasks. Nevertheless, the effect size estimate indicates that this significant is quite small." (p. 390). Their focus was more on global ratings, and, as they note, "results may be different if examined in terms of specific scales, e.g. grammar, pronunciation, etc." (p. 390). It is possible that different cultural groups focus on different aspects when evaluating the quality of learners' speech.

While Myford's (1996) study explored the relationship between number of foreign languages spoken and rater behavior, we are unaware of any study that investigates the possible effects that familiarity with the test taker's native language can have on raters' evaluations of their performance. Wiggleworth (1993) suggest the possibility of determining "whether there is any significant interaction between a rater

and a specific subgroup of candidates, e.g., candidates from particular language backgrounds." (p. 319). In a subsequent paper (Wigglesworth, 1994), she suggested, "Further investigations may provide insights into whether particular tasks, or even specific criteria, are biased in relation to particular subsets of the population to whom the test is administered. It is possible that people from a particular language or cultural background interact with specific tasks or criteria in a biased way." (p. 89). With her colleague (Chalhoub-Deville & Wigglesworth, 2005), she explored possible bias by raters from specific countries. Wigglesworth's research highlights the interest both in bias due to rater's nationalities and bias depending on the test taker's nationality or language background.

In our paper, we seek to better understand the connection between rater and learner language backgrounds. We focus both on the native and second languages of the raters and look for connections between rater language backgound and test taker native language. Specifically, our questions are as follows:

Are there significant differences between raters of an English EI test based on the native and second language backgrounds of these speakers?

If there are differences, do these differences vary according to the native language background of the English language learners who take the test?

Elicited imitation (EI) has been used for decades to evaluate the development of oral language skills studies of normal native language development (Ervin-Tripp, 1964; Menyuk, 1963; Keller-Cohen, 1981) abnormal language development (Menyuk, 1964; Berry, 1976; Lahey, Launer, and Schiff-Myers, 1983) and second language development (Naiman, 1974; Hamayan, Saegert, and Larudee, 1977). In recent years there has been a

resurgence of interest in its use for the examination of second language speaking development (Vinther 2002; Chaudron, Prior, and Kozok, 2005; Erlam 2006, Jessop, Suzuki, & Tomita, 2007). Erlam (2006) and Ellis (2005) have used EI as a measure of implicit L2 knowledge.

In an EI test, a speaker hears a sentence and then repeats the sentence as closely to the original as they are able. The process is the repeated for a series of sentences until the test is completed. Sentence reproductions are recorded for later rating. Bley-Vroman & Chaudron (1994) observe, "We regard it as premature to view elicited imitation as a proven method for inferring learner competence, because a considerable amount of research needs to be conducted to understand how performance under imitation conditions compares with other methods and with learners' underlying knowledge" (p. 245). However, with regards to the psychometric use of EI they claim that, "The more you know of a foreign language, the better you can imitate the sentences of the language. Thus, EI is a reasonable measure of global proficiency" (p. 247).

While the validity of EI as a measure of L2 proficiency or even implicit L2 knowledge can be debated, such debate goes beyond the scope of this paper. The purpose of this paper is to address possible rater bias, in particular when rating EI performance. For more comprehensive discussions of the merits of EI, see Gallimore and Tharp (1981), Lust, Chien, and Flynn (1987), Bley-Vroman and Chaudron (1994), Graham et al. (2008), Vinther (2002), Erlam (2006), and Ellis (2005).

One advantage of using the version of EI we have selected is that the rating criteria are relatively objective and scoring straightforward. We also hypothesize that, since students are all asked to repeat the same sentences, most of the variation in rating is

likely to be due to pronunciation rather than issues such as sociolinguistic and pragmatic mistakes, distracting grammatical errors, or idiosyncratic habits, etc. that might distract the rater. We recognize that this limits the generalizability of our study, but feel this is a starting point to determine whether raters of different language backgrounds are biased in a simple, relatively objective rating task that involves only discerning the sentences test takers produce and determining whether how well these sentences match native models.

In summary, in this paper we seek to determine possible rater bias due to the language backgrounds of both the raters and the test takers. Our analysis of bias involves data from elicited imitation, a test that requires raters to engage what should be a fairly objective task—mapping test taker productions to native models. We hope to determine whether familiarity with the test taker's native language biases the raters to give better or poorer ratings. Greater familiarity could allow raters to more easily discern test takers' production and therefore assign higher scores. On the other hand, it is also possible that raters could be stricter with speakers of languages they are familiar with, since their standard may be different for these speakers.

Methodology

To evaluate the possible relationships between rater attributes and test taker attributes (specifically, language background), we had twenty raters assess the performance of fifty test takers (learners of English as a second language) on ten elicited imitation items selected from a larger body of items from a more comprehensive test.

Test Takers

Test takers were students attending the English Language Center (ELC) at Brigham Young University in Provo, Utah. They included native speakers of Japanese, Korean,

Chinese, Spanish, or Portuguese learning English as a second language (ESL). Ten speakers of each language were chosen out of 760 total test takers using random stratified sampling in order to have two speakers from each of the five levels (Novice to Advanced) at the ELC for each language. Table 1 depicts the distribution of test takers selected for our ratings and the number of students originally tested in each language.

Raters

Twenty total native and non-native speakers of Japanese, Korean, Mandarin, Spanish, and Portuguese participated as raters in this study. Native speakers of these languages (n=10) were all highly proficient speakers of English as a second language. To control for the English proficiency level of these raters, only international students at Brigham Young University who were in their junior year or above were selected. Students who grew up speaking both English and the language listed as their native language were not qualified as raters because their English proficiency would not be comparable to the other non-native English-speaker raters. Non-native speakers of the five languages (all native speakers of English, n=10) had similar second-language learning experiences: they had served as missionaries for the Church of Jesus Christ of Latter-day Saints in areas where the second language was spoken by locals natively. Before they went abroad as missionaries, they had learned the languages in the U.S. at the Missionary Training Center for about 12 week. Subsequently, they lived for two years abroad and studied the languages on their own (i.e., as untutored learners). Given that the tasks they had to accomplish while abroad were similar, their second language proficiency was also similar upon return. Their proficiency ranged between Intermediate-High and Advanced-Mid on the ACTFL OPI.[i] To control for rating experience, a factor

shown to influence ratings in past studies (Myford, 1996), we selected only raters with no experience scoring elicited imitation samples. Table 2 depicts the number of native and non-native raters for each language. Fourteen of the raters were male and six female. Although gender has not been found to have a consistent effect on rater bias (Eckes, 2005; Schaefer, 2008), we will take this into account in our later analysis. Raters age, another factor that might have some influence on rater severity (Eckes, 2008), was controlled by selecting students at similar points in their educational experiences (ages 22-27). Table 2 depicts the rater makeup in terms of language background.

Administration and Scoring

Our EI was administered in a computer lab to individual classes at the ELC in conjunction with placement exams. Students logged on to EI program and were presented with a brief explanation of the research and an informed consent form. Following this, audio and video instructions were presented describing the test, telling students that they would hear each sentence only once, and instructing them to repeat items verbatim. A demonstration item with a model correct response was then played. Following this, students were given one practice item. If they had difficulty performing the task, students were asked to raise their hands for assistance. Once any difficulties were resolved, students proceeded on with the test. Items were presented to the learners one at a time in random order via high quality audio headsets. Responses were recorded using microphones attached to headsets. So, for example, for each item students saw on the screen a text that said "Sentence Number #." They then heard the sentence read by the male or female voice, followed by a beep signaling the beginning of the recording process. A time bar appeared on the screen showing the amount of time left to repeat the sentence. The time allotted to repeat sentences varied between six seconds for the short sentences and 12 seconds for the longest sentences. Once recorded, the files were saved as wave files for later analysis. Students completed sixty items total for the test. All sixty test-taker responses were stored on a server in a database for later scoring and analyses.

Rating Process

Before the raters started rating, they participated in a thirty-minute training session to learn how to rate and each practiced rating a few example sentences with the trainer present until they felt prepared to start rating and the trainer felt they were prepared to begin rating on their own.

Raters gave 1 point for each syllable in a sentence the students repeated correctly and 0 point when students were unable to correctly repeat a syllable. Written instructions on how to rate were given to each rater. Thus, throughout the rating process, the raters were able to refer the instructions provided in the training as needed, or they could contact their trainer to ask any questions that arose while rating.

There is no standardized method in the literature for scoring EI items (Vinther 2002). Those interested in determining whether learners control specific morphological or syntactic features of the language have usually examined each repeated sentence for the presence or absence of the target features while ignoring other inaccuracies which may have occurred in the repetitions (Erlam 2006, Munnich, Flynn, Martohardjono, 1994). Those attempting to develop an indirect method of estimating global language proficiency have generally scored items on a scale of correctness varying from a two point scale (Henning,1983), to a three point scale (Radloff, 1992), to a five point scale (Chaudron, et al 2005), to a seven point scale Keller-Cohen (1981). Lonsdale, Dewey, McGhee, Johnson, and Hendrickson (2009) experimented with a variable scale in which each syllable was awarded one point for being correct or zero points for being incorrect or absent. Differences in correlations with scores on oral proficiency interviews, between tests scored by assigning one point for each syllable produced correctly and those scored

by our a four point scale method (Chaudron et al., 2005) were small and inconsistent. Given this finding and the fact that syllable accuracy is used to generate numeric scores in other studies, we selected the syllable scoring method.

Ten sentences were selected according to difficulty ratings. Item Response Theory analysis was used to determine difficulty level, based on results from a sample of over five-hundred test takers at the ELC. Items were selected such that there two sentences that were appropriate in terms of difficulty for each of the levels at the ELC. This was done to assure that there would be adequate variation test results to see a range of performance.

A computerized scoring tool selected sentences in random order from the database until all five-hundred sentences (ten sentences times fifty students) were scored by each rater. Scores were input back into the database using the same scoring program and were later retrieved for analysis. The raw agreement by syllable between judges was .92 and interrater reliability (Kohen's κ) was 0.83.

Analysis

For various logistical reasons (hiring restrictions, visa issues, etc.), three of the raters were unable to finish all of the ratings (though they finished more than 94% each). Therefore, rather than generating a total score for each individual by each rater, scores for ten individual EI items were averaged and these averages were used as the dependent variable in the analysis.

A mixed linear model, blocking on individual rater, was used to evaluate relationships between rater variables (native and second language and gender), test taker variables (native language, ELC level, and gender), and EI ratings (averages). Rater

native language and second language were collapsed into one variable (rater language) to simplify analysis.

Results

Through backwards elimination, rater gender and all two-way interactions except for rater language X test taker L1 and test taker L1 X test taker gender were eliminated from the mixed linear model due to their low levels of significance. The final model (see Table 3) indicated a significant main effect for rater language, but no significant interaction between rater language and student language. Estimated marginal means are given in tables 4-9. Post-hoc pair-wise comparisons (Bonferroni) showed that, aside from one difference between the raters who were second language speakers of Korean (K2) and the native language speakers of Portuguese (P1) and Japanese (J1). Overall, the K2 raters were more generous in their ratings than the P1 and J1 raters. Main effects were also found for test taker L1, test taker gender, and test taker level. Post-hoc pair-wise comparisons showed that the Portuguese test takers (P1) scored significant lower than every other group. The Koreans (K1) scored significantly lower than Chinese (C1) and Spanish (S1) test takers. Females out-performed males on the test, and learners consistently performed better as they moved up in level (i.e., post-hoc pair-wise comparisons showed significant differences between levels) until levels 4 and 5. There were no significant differences between these levels (in fact, there was a slight non-significant decrease in score from Level 4 to Level 5). Finally, there was a significant interaction between test taker gender and test taker L1. While Spanish, Portuguese, and Japanese females performed higher than males, an opposite pattern was seen for Chinese and Korean.

Discussion

The focus of this paper was to determine whether there were differences in EI ratings based on the native and second language backgrounds of the raters. In answer to this question, we found that rater behavior did differ by language background, but we found no connection between these language backgrounds and the native languages of the test takers. In other words, there was no pattern of bias by which speakers of a particular language tended to favor or disfavor native speakers of that language in any way, regardless of their familiarity with typical difficulties English language learners/test takers have that may be common to native speakers of their language. One might expect, for example, that native or second language speakers of Spanish would be familiar with typical errors made by native Spanish speakers who are learning English; consequently, they might either understand these speakers better and assign higher ratings, or they might be stricter on these speakers, holding a higher expectation based on their own language learning experiences. Hinting at such a pattern, Myford (1996) found a tendency toward severity for individuals who spoke additional languages—the more languages they spoke, the stricter they were in their ratings of English language learners' writing. Again, we found no such pattern—raters were neither more severe nor more lenient as a whole, based on their language backgrounds. This difference could be due to two facts: first, our data involved spoken language, whereas Myford's involved written; second, in our study we sought to connect specific rater language backgrounds with test taker backgrounds, whereas Myford simply counted the number of additional languages a rater knew.

The absence of language-based rater bias may be attributable to the very objective nature of the EI ratings performed in this test. Raters are asked not to rate the quality of speech, but to simply determine whether learners produced individual syllables in a given sentence. Aspect of language such as fluency, vocabulary knowledge, grammatical accuracy, pragmatic competence, etc. are not evaluated by the raters, and global ratings that might take these aspects of language into account are not part of the EI scoring. In short, the very objective nature of the rating task may have prevented rater bias from surfacing.

EI rating can sometimes involve somewhat more subjective methods of evaluation. For example, Iwashita (2006) used the following scoring method for her analysis of EI results: 0= silence, garbled and unintelligible repetition, or minimal repetition of less than half of the idea units; 1= about half of idea units represented in string but a lot of information in the original is left out; or the string doesn't in itself constitute an independent sentence with some meaning; 2= more than half of the idea units are represented and string is meaningful, but it has some slight changes in content which make the sentence inexact, incomplete, or ambiguous; 3= the original meaning is preserved but there are some changes in the form of the string which may introduce some ungrammaticalities (but meaning doesn't change); 4= exact repetition. Using a 5-point rating rubric, Chaudron et al. (2005) and Graham (2006) produced an EI item score ranging from 0 to 4 for each sentence. Students started with a perfect score of 4 for each item. One point was then taken off for each syllable that was missing, unintelligible, or added. Participant responses that were missing more than three syllables were given a score of 0. Points were not taken off for mispronounced

words unless (1) the participant used a completely different word than the word in the prompt or (2) the response (or a part of it) was unintelligible. It is possible that bias could emerge in these relatively objective ratings, but it could also be that EI speech production in general is too structured and uniform to allow for significant amounts of rater bias. If this is the case, then it is certainly one of the benefits of EI, contributing to higher levels of consistency in rating speech samples. Future research could explore possible bias in other methods of scoring EI. It is also necessary to evaluate bias based on rater and speaker language backgrounds in other more holistic methods of evaluating spoken language abilities and in tasks that involve more complex language samples and a variety of linguistic production.

The finding of significant differences between raters with varying L1 and L2 backgrounds could be an anomaly. Overall, the K2 raters were more generous than their peers, though only more significantly so in the case of J1 and P1 raters, who tended to be more severe than others. Given that there were only two raters per language, it is possible that our raters were just unusually sever or lenient, independent of their native language or second language backgrounds. In other words, without using additional raters, it is difficult to say for sure whether raters from different L1 and L2 backgrounds are consistently more severe or generous in their ratings than others. Further research into rater language backgrounds could help us better understand the results of our study.

Rater gender and student gender were not found to be related in our results. However, our sample showed a main effect for test taker gender and an interaction effect for gender and test taker native language. The Chinese and Korean males performed slightly better than the females, but in other languages females tended to out-perform

males. A gender effect has been seen in studies of L2 speaking abilities during study abroad (e.g., Brecht, Davidson, & Ginsberg, 1993), with males out-performing females. The interaction effect seen here may be due to the fact that our sample from each language background was so small (ten individuals per language). While this difference is interesting, it does not indicate any rater bias and will therefore not be further addressed here.

As expected, as learners move up in level, their scores on EI increase. There seems to be a bit of a ceiling effect (i.e., learners don't show significant improvement from Level 4 to Level 5), but this ceiling effect has not been found in any of our previous studies (Graham et al., 2008; Hendrickson et al., 2009; Lonsdale et al., 2009). For this reason, we might conclude that our sample of learners was not representative of these levels. Again, this is a possible topic for future research.

While gender, level, and rater L1 showed some interesting patterns in our study, the main variable of interest was rater language background. Our finding of no significant interaction between rater language background and test taker L1 suggests that test administrators may not need to be concerned about matches and mismatches between rater and test taker languages. In other word, it appears that students will be scored roughly the same, regardless of matches or mismatches between their L1 and the languages the raters speak. This makes it possible for raters to evaluate speakers from a variety of language backgrounds: Chinese raters in China could potentially evaluate not only Chinese learners of English, but also French or Spanish learners of English. Native English speakers with L2 backgrounds in Spanish or Portuguese could safely evaluate speakers of these languages as well as native speakers of Chinese and Korean.

Conclusion

Test rater bias has been a topic of great interest in recent years. As Chalhoub-Deville & Wigglesworth's (2005) study indicated, nationality *can* lead to some rater bias on tests of written language. In our study, there were minimal differences between raters based on their native and second language backgrounds, and there were no connections between rater language background and test taker background. Thus, it appears test raters are not bias toward or against native speakers of languages that they (the raters) speak. If such bias is not present, then raters can safely evaluate the test performance of learners from a broad range of L1 backgrounds. While this study is informative, it is still small in scale and limited to EI. Future research involving larger numbers of raters and test takers and a variety of language tests are necessary in order to draw broader conclusions generalizable to other test settings.

Footnotes

Works Cited

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

Berry, P.B. (1976). Elicited imitation of language: Some ESNS population characteristics. *Language and Speech*, 1, 350–362.

Bley-Vroman, R. and Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. *Research methodology in second language acquisition*, 7, 245–261.

Brecht, R., Davidson, D., & Ginsberg, R. (1993). *Predictors of foreign language gain during study abroad.* Washington, DC: National Foreign Language Center. (ERIC Document Reproduction Service No. ED360828). (Reprinted in *Second language acquisition in a study abroad context*, pp. 37-66, by B. Freed, Ed., 1995, Amsterdam: John Benjamins)

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

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Working Papers in Second Language Studies, 21*(3), 1-44.

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and english language speaking proficiency. research report. *World Englishes, 24*(3), 383-391.

Chaudron, C., Prior, M. and Kozok, U. (2005). Elicited imitation as an oral proficiency measure. Paper presented 14 World Congress of Applied Linguistics, Madison Wisconsin .

Du, Y., & Others, A. (1996). *Differential facet functioning detection in direct writing assessment*

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

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.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: a psychometric study. *Studies in Second Language Acquisition*, 27, 141-172.

Erlam, R. (2006). Elicited imitation as a measure of l2 implicit knowledge: An empirical validation study. *Applied Linguistics*, 27, 464–491.

Ervin-Tripp, S. (1964). Imitation and structural change in children's language. In E.H. Lenneberg (Ed.), New directions in the study of language. Cambridge, MA: M.I.T Press, 163-189.

Gallimore, R. and R.G. Tharp (1981). The interpretation of elicited sentence imitation in a standardized context. *Language Learning*, 31, 369–392.

Galloway, V. B. (1980). Perceptions of communicative efforts of american students of spanish. *Modern Language Journal, 64*(4), 428-33.

Graham, C. R. (2006). An analysis of elicited imitation as a technique for measuring oral language proficiency. In Y. Chen and Y. Leung (Eds.), *Selected Papers from the Fifteenth International Symposium on English Teaching* (pp. 57-67). Taipei, Taiwan: English Teachers' Association.

Graham, C. R., Lonsdale, D., Kennington, C., Johnson, A. and McGhee, J. (2008). Elicited imitation as an oral proficiency measure with ASR scoring. Proceedings of LREC 2008.

Hamayan, E., J. Saegert, and P. Larudee (1977). Elicited imitation in second language learners. *Language and Speech,* 20, 86–97.

Hendrickson, Ross, Meghan Eckerson, Aaron Johnson and Jeremiah McGhee (2009). What makes an item difficult? A syntactic, lexical, and morphological study of Elicited Imitation test items. Proceedings of Second Language Research Forum 2008 (SLRF) [Forthcoming]

Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly, 3*(2), 151-169.

Jessop, L., Suzuki, W. and Tomita, Y. (2007). Elicited imitation in second language acquisition research, *The Canadian Modern Language Review*, 64, 1, 215-220.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring japanese second language writing performance. *Language Testing, 19*(1), 3-31.

Lahey, M., Launer, P.,and Schiff-Myers, N. (1983). Prediction of production: elicited imitation and spontaneous speech productions of language disordered children. *Applied Psycholinguistics*, 14, 317–343.

Lonsdale, D. Dewey, D. P., McGhee, J. Johnson, A., Hendrickson, R. (2009). Methods of Scoring Elicited Imitation Items: An Empirical Study. Paper presented at the annual conference of the American Association for Applied Linguistics, March 22, Denver, Colorado.

Lumley, T. and McNamara, T.F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing 12*, 54 – 71.

Lust, B., Chien, Y., and Flynn, S. (1987). What children know: methods for the study of first language acquisition. *Studies in the acquisition of anaphora*, 2, 271–356.

Menyuk, Paula (1963). A preliminary evaluation of grammatical capacity in children. *Journal of Verbal Learning and Verbal Behavior*, 2, 429–439.

Menyuk, Paula (1964). Comparison of grammar of children with functionally deviant and normal speech. *Journal of Speech and Hearing Research*, 7, 109–121.

Naiman, Neil (1974). The use of elicited imitation in second language acquisition research. *Working Papers on Bilingualism*, 2, 1–37.

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

Tanaka, M., Hajikano, A. and Tsubone, Y. (1998) Daini gengo to shite no nihongo ni okeru sakubun hyooka : ‘Ii’ sakubun no kettei yooin [Evaluation criteria for writing by non-native speakers of Japanese : factors affecting the evaluation of

‘good ’ writing]. *Nihongo Kyooiku* [Journal of Japanese Language Teaching] *99*, 60 – 71.

Vinther, T. (2002). Elicited imitation: a brief review. *International Journal of Applied Linguistics*, 12, 54–73

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

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.

Table 1

Number of students from rated by language and level and number of students in original test sample.

	Level 1	Level 2	Level 3	Level 4	Level 5	Total	Original Sample
Japanese	2	2	2	2	2	10	62
Korean	2	2	2	2	2	10	196
Chinese	2	2	2	2	2	10	89
Spanish	2	2	2	2	2	10	347
Portuguese	2	2	2	2	2	10	66
Total	10	10	10	10	10	50	760

Table 2

Number of Raters Per Language.

	Japanese	Korean	Mandarin	Spanish	Portuguese	Total
Native Speakers	2	2	2	2	2	10
Non-Native Speakers	2	2	2	2	2	10
Total	4	4	4	4	4	20

Table 3

Mixed Linear Model Results

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	17.2	1912.34	.000
Student L1	4	772.3	8.52	.000
Student Gender	1	897.5	59.67	.000
Student Level	4	795.7	117.63	.000
Rater Language Background	9	907.1	3.11	.001
Student L1 * Rater Language Background	36	907.5	.44	.999
Student L1 * Student Gender	4	839.9	19.48	.000

Table 4

Estimated Marginal Means by Test Taker Native Language (L1)

Test Taker L1	Mean	Std. Error	df	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Chinese	.646	.018	36.0	.611	.682
Japanese	.635	.018	37.1	.599	.671
Korean	.600	.017	36.7	.565	.635
Portuguese	.574	.018	38.8	.538	.610
Spanish	.654	.018	39.6	.617	.690

Table 5

Estimates by Gender

Gender	Mean	Std. Error	df	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
F	.658	.015	20.442	.627	.689
M	.586	.015	21.279	.555	.617

Table 6

Estimates by ELC Level

Level	Mean	Std. Error	df	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1	.457	.018	40.497	.421	.493
2	.515	.018	38.853	.479	.551
3	.624	.017	32.084	.589	.658
4	.760	.019	43.945	.722	.797
5	.753	.018	39.658	.717	.789

Table 7

Estimates by Rater Language Background

Rater Language	Mean	Std. Error	df	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
C1	.624	.019	48.837	.587	.661
C2	.643	.018	48.727	.606	.680
J1	.589	.019	52.331	.552	.627
J2	.636	.018	48.787	.599	.673
K1	.621	.018	48.615	.584	.658
K2	.656	.018	48.690	.619	.693
P1	.585	.018	48.716	.548	.622
P2	.611	.019	52.615	.573	.649
S1	.618	.018	48.629	.581	.655
S2	.634	.019	48.846	.596	.671

Table 8

Estimates for Test Taker L1 X Rater Language Background

L1	RaterLang	Mean	Std. Error	df	95% Confidence Interval Lower Bound	Upper Bound
Chinese	C1	.653	.032	308.170	.591	.716
	C2	.661	.032	310.868	.599	.723
	J1	.632	.032	305.157	.569	.695
	J2	.654	.032	310.868	.592	.717
	K1	.649	.032	305.729	.586	.712
	K2	.686	.032	305.157	.623	.749
	P1	.603	.032	305.668	.540	.666
	P2	.623	.033	344.752	.558	.689
	S1	.639	.032	306.441	.577	.702
	S2	.660	.032	309.902	.598	.723
Japanese	C1	.646	.032	309.384	.584	.709
	C2	.674	.032	309.384	.612	.737
	J1	.570	.033	348.072	.504	.635
	J2	.658	.032	309.384	.595	.720
	K1	.626	.032	308.617	.564	.688
	K2	.652	.032	309.514	.589	.715
	P1	.583	.032	309.514	.520	.646
	P2	.632	.033	348.583	.567	.697
	S1	.637	.032	309.125	.575	.700
	S2	.675	.032	309.384	.612	.737
Korean	C1	.599	.032	315.626	.537	.661
	C2	.607	.032	314.982	.544	.669
	J1	.577	.034	378.274	.511	.644
	J2	.632	.032	314.982	.570	.694
	K1	.618	.032	315.531	.555	.680
	K2	.639	.032	316.149	.576	.701
	P1	.571	.032	315.171	.509	.633

L1	RaterLang	Mean	Std. Error	df	95% Confidence Interval Lower Bound	Upper Bound
	P2	.545	.033	336.728	.481	.609
	S1	.595	.032	314.806	.533	.657
	S2	.617	.032	315.626	.555	.679
Portuguese	C1	.564	.032	313.612	.501	.627
	C2	.579	.032	312.840	.516	.642
	J1	.561	.032	313.612	.498	.623
	J2	.579	.032	313.612	.517	.642
	K1	.554	.032	313.612	.491	.617
	K2	.605	.032	313.612	.542	.668
	P1	.540	.032	313.612	.477	.603
	P2	.618	.033	331.333	.554	.681
	S1	.569	.032	313.612	.506	.631
	S2	.568	.032	313.612	.506	.631
Spanish	C1	.659	.032	312.985	.596	.722
	C2	.695	.032	312.985	.632	.758
	J1	.608	.033	350.631	.542	.673
	J2	.655	.032	312.985	.592	.718
	K1	.657	.032	312.103	.594	.719
	K2	.698	.032	311.411	.635	.760
	P1	.630	.032	311.557	.567	.693
	P2	.638	.033	351.578	.572	.703
	S1	.650	.032	312.103	.587	.713
	S2	.648	.032	312.985	.585	.711

Table 9

Estimates for Student Native Language X Student Gender

L1	StudentGender	Mean	Std. Error	df	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Chinese	F	.632	.020	61.438	.592	.672
	M	.660	.021	68.557	.618	.702
Japanese	F	.719	.020	62.384	.679	.759
	M	.552	.021	68.836	.509	.594
Korean	F	.578	.021	72.686	.536	.620
	M	.622	.020	60.317	.582	.662
Portuguese	F	.650	.019	53.915	.611	.689
	M	.497	.024	93.816	.450	.544
Spanish	F	.709	.022	71.193	.666	.752
	M	.598	.021	68.276	.557	.640

[i] See Dewey and Clifford (2010) for a more detailed description of the language learning experiences and proficiency levels of these returned missionaries.