2016-03-01

The Effect of Prompt Accent on Elicited Imitation Assessments in English as a Second Language

Jacob Garlin Barrows
Brigham Young University - Provo

The Effect of Prompt Accent on Elicited Imitation Assessments in English as a Second Language

Jacob Garlin Barrows

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

Master of Arts

Troy Cox, Chair
Wendy Baker-Smemoe
Dan P. Dewey

Department of Linguistics and English Language

Brigham Young University

March 2016

ABSTRACT

The Effect of Prompt Accent on Elicited Imitation Assessments
in English as Second Language

Jacob Garlin Barrows
Department of Linguistics and English Language, BYU
Master of Arts

Elicited imitation (EI) assessment has been shown to have value as an inexpensive method for low-stakes tests (Cox & Davies, 2012), but little has been reported on the effect L2 accent has on test-takers' ability to understand and process the test items they hear. Furthermore, no study has investigated the effect of accent on EI test face validity. This study examined how the accent of input audio files affected EI test difficulty as well as test-takers' perceptions of such an effect. To investigate, self-reports of students' exposure to different varieties of English were obtained from a pre-assessment survey. A 63-item EI test was then administered in which English language learners in the United States listened to test items in three varieties of English: American English, Australian English, and British English. A post-assessment survey was then administered to gather information regarding perceived difficulty of accented prompts. A many facet Rasch analysis found that accent affected item difficulty in an EI test with a separation reliability coefficient of .98—British English being the most difficult and American English the easiest. Survey results indicated that students perceived this increase in difficulty, and ANOVAs between the survey and test results indicated that student perceptions of an increase in difficulty aligned with reality. Specifically, accents that students were "Not at all Familiar" with resulted in significantly lower EI test scores than accents with which the students were familiar. These findings suggest that prompt accent should be carefully considered in EI test development.

Keywords: elicited imitation, language testing, accent

ACKNOWLEDGMENTS

I would like to thank my committee for their excellent feedback and support, especially Troy Cox for his mentoring and endless patience. I would also like to thank Judson Hart and the other employees of the English Language Center who facilitated this research as well as the other BYU students who volunteered their time to help. Finally, special thanks go to my wife, Sarah, for her support; her sister, Elizabeth, for her assistance with this research; and Sybil Lewis and David, Moira and Marjorie Barrows, who paved the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Globalization and modern communication and media technology are bringing new challenges to the field of language assessment. While students are increasingly likely to travel and study a major foreign language, they are also more likely to encounter new, challenging varieties of their target language. This is especially true of world languages—such as Spanish or English—which enjoy the privileged status of being studied and spoken internationally in many different contexts, including business, entertainment, and academics. English in particular has developed robust L1 and L2 varieties both regionally and nationally, such that a speaker or learner might only be exposed to one or two varieties. A student of Spanish in Europe, for example, might never have to communicate with a Mexican and might never be exposed to media from Mexico. How, then, would a speaking or listening assessment accurately measure that student's ability if the prompts contained a Mexican variety of Spanish? Likewise, an Indian speaker of English might never have the opportunity to speak with a native speaker of an Inner Circle variety of English. How fairly would a test designed with British, American, or Australian English assess that speaker's ability?

This raises potential difficulties for assessing listening and speaking, as those who design tests for an international audience cannot make broad assumptions concerning a learner's background with a given variety of the language. Care must be taken to assure that a listening or speaking assessment measures actual ability and is not affected by a learner's familiarity (or lack thereof) with the assessment variety. This problem is compounded by the mobility of many language students (who can come from practically any place or language background) and innovative assessments that can be administered anywhere in the world via the internet.

While this is a challenge for all types of language assessment, it is a particular challenge

for speaking and listening assessments, which typically require an interlocutor or audio prompts. Any audio prompt (or interlocutor's speech) is by necessity colored by the speaker's accent, which may add an additional layer of difficulty for those unfamiliar with that variety. As most interlocutors only speak a single variety, oral proficiency interviews risk putting some test-takers at a disadvantage; but even if an enterprising test designer included audio recordings from multiple varieties, they would still be faced with the impossible task of selecting the perfect cocktail of varieties that would fairly assess all test-takers.

To address this challenge, this research seeks to better understand the interaction between a learner's familiarity with regional varieties and their results on an elicited imitation (EI) speaking assessment. EI is a relatively new testing technique that has the potential to expedite the assessment process and reduce cost (Cox, Bown, & Burdis, 2015). The design of EI assessments is fairly simple: students listen to a number of audio prompts and attempt to repeat verbatim what they hear. Their repetition is recorded and later graded for accuracy. With this type of assessment, many students can complete the test simultaneously in a computer laboratory setting and receive rapid feedback; alternatively, they can complete the test from nearly any location using a web-based EI test. The time and cost benefits of EI testing, as well as the flexibility of test administration, make it an attractive alternative to traditional oral proficiency interviews (OPIs) in some low-stakes situations, and a well-designed EI test can accurately predict OPI outcomes (Cox, Bown, & Burdis, 2015).

Another challenge of EI testing is face validity (Graham, Lonsdale, Kennington, Johnson, and McGhee, 2008; Van Moere, 2012; Vinther, 2002; Moulton, 2012), meaning that some language testing professionals and test-takers have difficulty seeing how a sentence repetition task could measure language proficiency. Anecdotes and comments from students suggest that

this problem is compounded (i.e. students have less trust in test results) when test-takers listen to EI prompts in an accent they are less familiar with, though the extent of these perceptions have never been studied formally. This is important as previous research on the topic indicates that low face validity may hinder test-taker motivation and therefore test performance (Chan, Schmitt, DeShon, Clause, & Delbridge, 1997).

To improve the accuracy and validity of EI testing, this research aims to answer the following questions:

1. What effect does speaker accent have on EI test item difficulty?
2. What are students' perceptions of the effect of accent on EI test difficulty and to what extent are these perceptions accurate?

In order to answer these questions, a new EI test was created at BYU's English Language Center (ELC) to include recordings of American, Australian, and British speakers. This test was accompanied by two surveys: a pre-test survey, which gathered data on participants' experience with and exposure to different varieties of English, and a post-test survey, which gathered data on participants' perceptions of how accent affected test difficulty. To answer the first research question, the results of the test and pre-test survey were analyzed using a Rasch facets analysis. To answer the second research question, the results of the pre- and post-test survey, along with the test scores, were analyzed.

2. Review of Literature

This chapter begins by acknowledging previous research on the interaction between listening comprehension and accent. These studies have been theoretical trailblazers for the current study, and such a review will provide a) context and justification for the current study and b) a defensible basis for the current study's definition of accent and listening comprehension. Elicited imitation assessment will then be addressed and located in the aforementioned theoretical foundations.

2.1 Accent and Listening Comprehension

A method of directly measuring listening comprehension—among other receptive skills—has thus far eluded researchers; instead, they have had to devise means of measuring it indirectly (Derwing & Munro, 2009). While all of these methods are designed to measure listening comprehension, they are unique in what they actually measure and how their creators define—often implicitly—accent and listening comprehension. No previous study has investigated accent or listening comprehension in exactly the same way this study does. It is still useful, however, to consider previous research, as it sheds light on accent and listening comprehension generally and lays the groundwork for the theoretical foundations of this study.

In this review of literature, *ESL* (English as a Second Language) refers to the study of English in an English-speaking context (e.g. studying English in the US). *EFL* (English as a Foreign Language) refers to the study of English in a non-English context (e.g. studying English while living in Japan). *L1* listeners and speakers refers to those whose first language is English. *L2* listeners and speakers refers to those who are learning or have learned English as a second (or even third, fourth, etc.) language; this term is broad and may include both ESL and EFL learners. When discussing different L1 varieties of English, *international* varieties refer to L1 varieties

that do not belong to the listener's own country or, in the case of L2 listeners, L1 varieties of English other than those which have been learned or studied (e.g. Australian English for American listeners or for ESL students in America). *Regional* varieties refer to those that are sub-categories of international varieties (e.g. the English of the American South).

2.1.1 Effect of accent on listening comprehension. Of the studies investigating the interaction between listening comprehension and accent, nearly all have found that an unfamiliar accent hinders listening. This has been found with different types of accent, including L2-accented English (e.g. Varonis & Gass, 1982; Gass & Varonis, 1984; Anderson-Hsieh & Kohler, 1988; Clarke & Garrett, 2004), regional and international accents for L1 listeners (Adank & McQueen, 2007; Adank, Evans, Stuart-Smith, & Scott, 2009; Major, Fitzmaurice, Bunta, & Balasubramanian, 2005; Floccia, Goslin, Girard, & Konopczynski, 2006), regional and international accents for L2 listeners (Ockey & French, 2014; Major *et al.,* 2005), L1 ethnic accents (Major *et al.,* 2005), and even artificial accents (Wingstedt & Schulman, 1984; Maye, Aslin, & Tanenhaus, 2008).

There are two notable exceptions, however, both of which investigate EFL listeners. First, Abeywickrama (2013) found no effect of Chinese, Korean, and Sri Lankan accented English on the scores of a Test of English as a Foreign Language (TOEFL) style listening test that was administered to students of English residing in Korea, Brazil, and Sri Lanka. This test included eight listening passages read by two groups of speakers: the first group acted as control and included four Americans; the second group included one Korean, one Sri Lankan, and two Chinese speakers. Each listening passage was followed by three to four multiple choice questions. A likely cause for the mismatch between these results and those of similar studies is how Abeywickrama (2013) conceptualizes accent. Accent is not defined in this study, and it is

implied that the speakers are representative of a particular accent by virtue of their nationality. The speakers' performance on two tests of general speaking proficiency—the Test of Spoken English (TSE) and the Spoken English Proficiency Assessment Kit (SPEAK)—was the only indicator of their strength of accent, and even though they all scored between 50 and 60 out of a maximum of 60, Abeywickrama (2013) took this as evidence that their speech was distinctly nonnative. These issues indicate that when researching any effects of accent, accent must clearly be defined in terms of what it is and how strong it is.

The second study investigated whether a shared native language aided listening comprehension when EFL students listened to nonnative English speakers on a TOEFL-like listening test (Major, Fitzmaurice, Bunta, and Balasubramanian, 2002). They found that the Spanish speakers in their study performed better when listening to Spanish-accented English (compared to American-accented English) and that the Chinese speakers performed worse when listening to Chinese-accented English. To explain these differences, the authors suggest that strength of accent may have played a role. The Spanish speakers who performed the recordings had much lighter accents when compared to the Chinese and Japanese speakers, as established by 76 judges who rated the accents on a 5-point scale. Evidence of the effect of accent could be seen in a closer examination of the data, which showed that most speakers performed worse with Chinese-accented English, and that Chinese, Japanese, and American participants all scored just as well with Spanish-accented speech as they did with American-accented speech. Beyond strength of accent, Major *et al.* (2002) pointed to phonological reasons, highlighting the fact that Chinese, Japanese, and Spanish are all similar in terms of rhythmic timing—Spanish and Mandarin Chinese are both syllable-timed languages and Japanese is a mora-timed language—contrasting with English, which is stress-timed. Transfer of this phonological feature from

Spanish to English, it is implied, may have made the speech more familiar to Chinese and Japanese speakers. This last claim is supported by Anderson-Hsieh, Johnson, and Kohler (1992), who found that prosody variation affects listening accuracy more than segmental variation. One factor that is taken for granted by Major *et al.* (2002) is how familiar the speakers were with English that is accented by their native language. It is assumed that Chinese speakers were accustomed to hearing Chinese-accented English or that phonological transfer would make them naturally familiar with it. This was perhaps a reasonable assumption at the time, but the curious results of their study suggest that accent familiarity should be more explicitly examined and measured.

It is worth noting that these two studies investigated the effect of L2 accent on EFL students who were not residing in an English speaking country (Abeywickrama, 2013, asserts that Sri Lanka is an ESL environment—implying that English is broadly spoken there—but whether the average Sri Lankan student's daily exposure to English differs from typical EFL environments is debatable). There is perhaps a difference between ESL and EFL listeners that future research should investigate.

Despite these two exceptions, however, the overwhelming consensus from the other studies previously noted indicate that, generally speaking, unfamiliar accents impair listening comprehension.

2.1.2 Accent. Research attempting to examine accent must apply a clear and defensible definition of accent. This is particularly true of the current study, which considers strength of accent and must therefore measure it. Clearly defining accent is a difficult task, however, as evidenced by the number of studies that fail to do so (e.g. Clarke & Garrett, 2004; Gass & Varonis, 1984; Abeywickrama, 2013).

One vein of thought is that accent is an attribute of speech that varies according to geography, native language (when nonnative speech is examined), ethnicity, or individual speaker (e.g. Clarke & Garrett, 2004; Anderson-Hsieh & Kohler, 1988; Abeywickrama, 2013). Those who take this perspective often view accent as a unidimensional construct for which a speaker's provenance is the only variable. In these studies, a speaker's identity or linguistic background is usually sufficient evidence of accent, though in some cases expert judgment (Adank *et al.*, 2009) or phonetic analysis is conducted to empirically delineate geographical accent boundaries (Adank & McQueen, 2007).

For the current study, this perspective on accent presents a number of challenges, including questions about how a national accent is defined, whether a national "standard" is representative of what the students are likely to have encountered, or whether such a standard is actually spoken in Australia, the US, and the UK—the three countries whose accents are considered here. This is particularly true of the UK, whose traditional standard, Received Pronunciation, has been on the decline for several years in favor of more regional varieties (Mugglestone, 2007). More importantly, though, this perspective is only focused on the speaker and completely ignores the listener's experience.

In listening research, a more appropriate definition of accent should account for the complex nature of language as a shared experience between two interlocutors. Thus, Derwing and Munro (2009) take a different approach and define accent as an attribute of listeners' perception rather than as an attribute of speech itself. They define the construct of accent as having three related, yet partially independent, dimensions: *accentedness*, in their view, is how different one variety sounds from the listener's local variety; *comprehensibility* is how difficult a listener believes a variety is to understand; and *intelligibility* is how much is actually understood. The first two are

based on listener judgments while the third is based on listener performance.

The partial independence of these dimensions was first verified by Munro and Derwing (1995a), who had a group of native English listeners complete a transcription task for speech samples recorded in Mandarin-accented English (this served as a measure of intelligibility) and then rate the samples for accentedness and comprehensibility. For most listeners, they found a positive correlation between *accentedness-comprehension* and a negative correlation between *comprehension-intelligibility*, but the strength of this correlation varied widely between listeners (ranging from 0.41 to 0.82 and -0.44 to -0.90, respectively); and they found no correlation between *accent-intelligibility* for most listeners. A closer look at the data shed further light on the relationship between accent and intelligibility: they noted that "the listeners apparently perceived a wide range of accentedness in stimuli that were nonetheless perfectly transcribed," suggesting that even strongly accented speech can be completely intelligible. These findings were later confirmed and elaborated on in other experiments (e.g. Munro & Derwing, 1995b; Derwing & Munro, 1997).

In the present study, EI test performance corresponds to the above definition of *intelligibility* (meaning that EI repetitions reflect what a student understands), test-takers' perceptions of how difficult the accents are to understand correspond to *comprehensibility*, and strength of accent corresponds to the degree of *accentedness* as determined by a panel of judges.

Strength of accent. Defining accentedness as a scalar dimension requires researchers to obtain a measurement of it. Derwing and Munro, who pioneered this definition of accent, obtained this measurement by asking the listeners in their experiments to rate the audio samples on a 9-point scale (Munro & Derwing, 1995a; Munro & Derwing, 1995b; Derwing & Munro, 1997). These listeners participated in every aspect of the study—both the accent judgment tasks

and the listening comprehension tasks—and were native speakers of English.

This contrasts with Ockey and French (2014), who, when conducting a similar study, only required test subjects to participate in the listening comprehension task. A separate panel of judges were used to measure accentedness. Since the listeners in their experiment were nonnative speakers of English at a wide range of proficiency levels, they determined that accentedness would be better measured by native and highly proficient nonnative speakers. Specifically, their research aimed to compare Standard American English (SAE) with other varieties, so all judges were residents of the US. Similar strength of accent scales and panels of judges have also been used in related studies (Major *et al.*, 2002; Major *et al.*, 2005).

For practical reasons, it is useful to be able to say that a speaker who comes from Australia speaks Australian English, but place of origin is insufficient evidence of how a person's speech is actually perceived by listeners. In the current study, a speaker's origin was used only as a point of departure; a strength of accent survey adapted from Ockey and French's (2014) survey is subsequently used to measure accentedness according to Derwing and Munro's (1995a) definition.

2.1.3 Accent and test face validity. Elicited imitation tests have typically suffered from low face validity (*validity* refers to how well a test measures what is intended; *face validity* refers to how well a test *appears* to measure what is intended be valid) (Graham *et al.*, 2008; Van Moere, 2012; Vinther, 2002; Moulton, 2012), which may be exacerbated when audio prompts contain unfamiliar accent. Comments from EI test-takers suggest that their trust in the test results sometimes decreases when accented audio prompts, though this has never been explicitly researched. Previous research on face validity in testing indicates that low face validity may lead to lower motivation on the part of test-takers, which in turn may reduce test performance (Chan

et al., 1997).

2.1.4 Listening comprehension. Another thing that must be defined and measured is listening comprehension. On the surface this appears to be fairly straightforward, as listening comprehension seems so obvious that it hardly requires an explanation. A closer look at past studies on accent and listening comprehension, however, reveals that previous research on the subject has measured a number of different things and that the term *listening comprehension* is perhaps too broad to be used without qualification.

In some studies, listening comprehension referred to performance on an academic test. Some researchers designed a listening test that was similar to the listening portion of the TOEFL—where students listen to a series of audio passages and answer three to four multiple choice questions for each one (e.g. Anderson-Hsieh & Kohler, 1988; Abeywickrama, 2013)—while Ockey and French (2014) used actual TOEFL data.

In other studies, listening comprehension referred to how accurately a person could respond to stimulus. Gass and Varonis (1984) created a written imitation test where subjects listened to and then transcribed a sentence. These transcriptions were scored for accuracy to determine what was comprehended. A similar method was also used in later studies (e.g. Munro & Derwing, 1995a, Munro & Derwing, 1995b; Derwing & Munro, 1997). Maye *et al.* (2008) used a lexical decision task; though subjects were required to respond within a two-second window, only accuracy was considered in the analysis.

Other studies have measured response time as well as accuracy, suggesting that listening comprehension is a combination of accuracy and processing speed. Adank *et al.* (2009) recorded response times and accuracy for a true-false decision task, Clarke and Garrett (2004) for a cross-modal matching task where participants considered an audio sample and a visual probe word,

Floccia *et al.* (2006) for a lexical decision task, and Adank and McQueen (2007) for an animacy decision task.

Derwing & Munro's (2009) definition of accent might also be applied to listening comprehension. Viewed from this perspective, some studies have investigated performance, measuring the effect of accent on *intelligibility*; others have investigated indicators of cognitive processing, such as response time, to measure the effect of accent on *comprehensibility*. The distinction between these types of listening has been confirmed in some studies (Adank *et al.*, 2009; Munro & Derwing, 1995a) and disputed in others (Sommers, Nygaard, & Pisoni, 1994; Clarke & Garrett, 2004). Adank *et al.* (2009), for example, found that an increase in response time does not necessarily lead to a decrease in response accuracy. Findings in Sommers, Nygaard, and Pisoni (1994), however, found evidence contradicting the previously mentioned research and suggest that an increase in cognitive processing is directly related to a decrease in performance. Though some research supports the independence of *comprehensibility* and *intelligibility*, further research on the subject is needed.

While EI testing is designed to measure performance—and therefore intelligibility—it is unique in that it is heavily dependent on working memory to hold the input while it is parsed into meaningful chunks. It is conceivable, therefore, that intelligibility in an EI task is more dependent on comprehensibility than in other listening tasks and that an increase in processing cost—brought on by a decrease in comprehensibility—might affect intelligibility in EI more than in other types of language testing. Listening in EI, therefore, must be investigated specifically since the findings of studies of other types of listening cannot be applied directly.

2.2 Elicited imitation

In elicited imitation assessment, test-takers are provided a prompt that they must attempt

to repeat verbatim. In many assessments, including the one used in this study, the test administration is automated, with test-takers listening to audio recordings and repeating the utterances into a microphone. The recorded utterances are later scored by marking errors in the repetition (e.g. the number of words or syllables incorrectly repeated or omitted).

The nature of EI test administration and scoring offers a number of advantages (technology permitting), among which time and cost effectiveness are most notable. When it is automated, an EI test can be administered to several students at the same time under the supervision of an administrator, even one who is not highly trained in the task (Graham *et al.*, 2008). Test scoring can also be quick and objective (Matsushita & Lonsdale, 2012) and, when scored by humans, doesn't require extensive training or even a native speaker to rate accuracy (Lonsdale & Millard, 2014). Future developments in automated speech recognition (ASR) technology may even further reduce the cost and increase the reliability of EI assessment (Cox, Bown, & Burdis, 2015).

These advantages are particularly noteworthy when compared to the cost of oral proficiency interviews (OPIs), which require one-on-one contact with a trained interviewer. Furthermore, OPIs are often double-rated for improved reliability. While OPIs may still be better suited to certain contexts, EI tests can be cost-effective alternatives in many low-stakes testing situations (Cox & Davies, 2012, provides a thorough comparison of the two assessments).

EI assessment has not been without its challenges, however. Two questions have often been asked of EI that are especially relevant to this research: 1) Does EI depend on comprehension or is it based purely on memory? And 2) what exactly is assessed—listening or speaking ability?

A study by Okura and Lonsdale (2012) set out to answer the first question. They created

two parallel EI tests—one with English sentences and the other with random nonce syllables—and administered them to a diverse group of ESL students. The first test resembled a typical EI test, but the second was essentially a test of working memory since there was no linguistic structure underlying the input. A significant correlation between performance on the two tests would suggest that they test the same thing (i.e. memory), but this was not found ($r = .249$, $n = 40$, $p = 1.21$). Though EI testing obviously involves working memory, it nonetheless hinges on something else: the ability to decode and comprehend linguistic input. This latter point was demonstrated in this same study, where a significant positive correlation was found between scores of the English EI test and the students' level in their ESL institution.

As for the second question, Vinther (2002) provides a thorough overview of the subject and points out that although EI directly measures utterances, those utterances could not be repeated without being understood. It is also true that subjects' capacity to repeat what they comprehend is limited by their speaking ability. It is therefore difficult to say, according to Vinther (2002), which ability carries more weight in an EI task: listening or speaking.

More recent research, however, has shed more light on the matter. A study by Cox and Davies (2012) comparing an automatically scored EI test with a number of other proficiency measures (i.e. a speaking proficiency interview, a writing placement exam, and a computer adapted exam of listening, reading, and grammar) found that although there were correlations between EI and the other assessments, the computer adapted listening exam and the EI test had the highest correlation ($r = .74$). This suggests, they say, that strong listening skills enable students to utter more accurate repetitions.

These questions are particularly relevant given the research on listening and comprehension that was noted above. It has been demonstrated that the process of imitation does

indeed depend on comprehension and that whether or not it is speaking that is assessed in EI, comprehension is the gatekeeper for imitated speech. It has also been demonstrated that accent—and familiarity therewith—can affect listening comprehension. It is a reasonable conclusion, then, that accent in EI audio prompts may affect EI test results. What remains to be seen is the degree to which accent affects EI test results and the role that familiarity with accent plays in this effect.

3. Methodology

This research aims to answer the following questions:

1. What effect does familiarity with regional accent have on EI test scores?
2. What are students' perceptions of the effect of accent on EI test difficulty and to what extent are these perceptions accurate?

In order to answer these questions, three instruments were created: an EI test with recordings of American, Australian, and British speakers, and two surveys—a pre-test survey gathering data on participants' experience with and exposure to different varieties of English, and a post-test survey gathering data on participants' perceptions of how accent affected their test results. To answer the first research question, the results of the test and pre-test survey were analyzed using a Many Facet Rasch Measurement (MFRM). To answer the second research question, the results of the pre- and post-test survey along with the test scores were analyzed.

3.1 The EI Test

This study required the development of a new EI test. A previous test that had already been validated provided the framework for the new test, including the text for most of the items; this test was designed to test general proficiency. Some new items were created in order to be validated during this round of testing. New audio recordings of the target accents needed to be obtained for all of the items, and it was essential to find speakers whose accents were of comparable strength. This was to ensure that the results were based on accent variety, not accent strength.

3.1.1 Selection of speakers. The speakers for these recordings were selected from nine volunteer undergraduate students at Brigham Young University. Two were American, two were Australian, and five were from the UK; each of the non-American volunteers had resided in their

country of origin until coming to the university between six months and three years before the recordings. Eight of the volunteers were between the ages of 18 and 28, but one British volunteer was 45. One of the Americans and three of the British volunteers were male while the other five were female. Each of the volunteers read the full list of sentences in a sound recording booth equipped with Shure microphones (model KSM32) to ensure the audio quality was high.

Samples of these recordings were then inserted into a strength of accent survey. The samples that were selected included two sentences per item difficulty level (intermediate, advanced, and superior) for each speaker, for a total of 54 sentences. Each item of the survey asked participants to "identify how different the English sounds from [their] local variety", after which they were presented with the audio clip and a 7-point Likert scale ranging from "not at all different" to "very different". The items were presented in random order, with only one item presented at a time. Survey participants did not have the option to go back and change their responses. A sample question from this survey can be found in Figure 2. This strength of accent scale and the question used to elicit a response was adapted from previous research (Major *et al.*, 2002; Major *et al.*, 2005; Ockey & French, 2014). Though the main purpose of this survey was to select speakers who had relatively similar accent strength according to listeners of other countries, it also had the added benefit of ensuring that the Australian and British volunteers still retained their home accents despite their stay in the US.

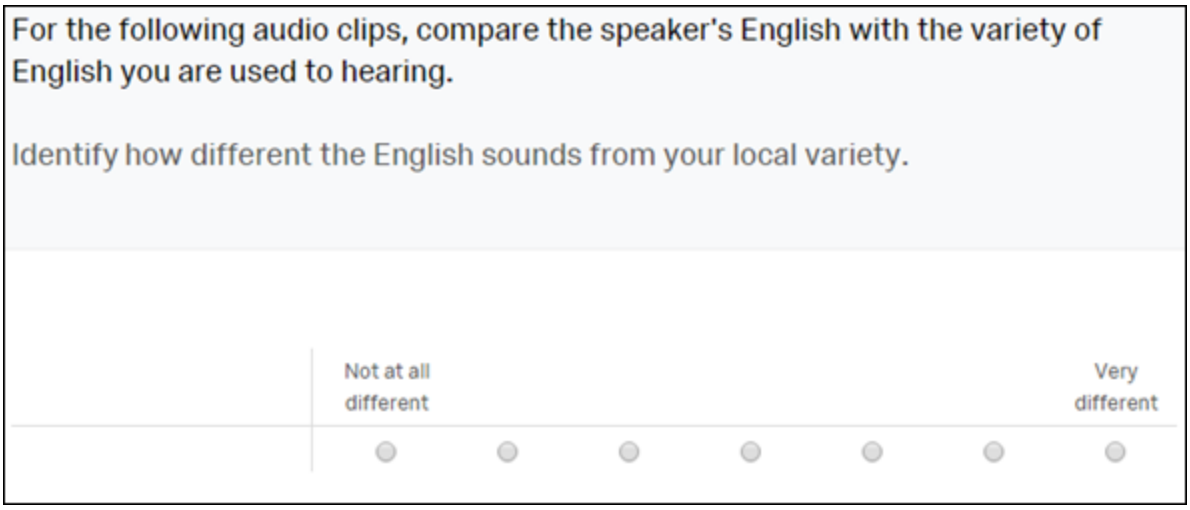

Figure 1. Sample Question from the Strength of Accent Survey

The survey was distributed online via Qualtrics. This distribution platform was selected because it provided easy access to respondents from the US, Australia, and the UK, which was necessary to measure strength of accent from each perspective. The survey was administered to 126 participants—42 from each country—each of whom was a native speaker of English, between 18-30 years of age, and had lived in their country of origin for the last ten or more years. The geographical distribution of survey participants within their respective countries was estimated by determining the geolocation of each participant's Internet Protocol (IP) address. This data can be found in Figures 2, 3, and 4.

Figure 2. Geolocations of the 42 Australian Participants in the Strength of Accent Survey

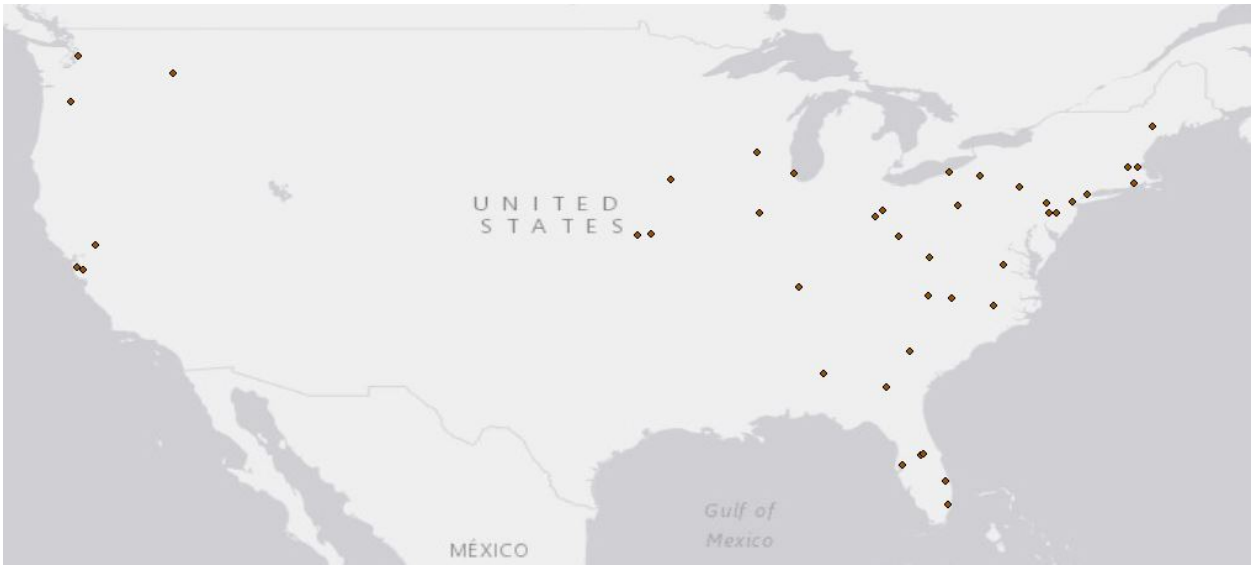

Figure 3. Geolocations of the 42 American Participants in the Strength of Accent Survey

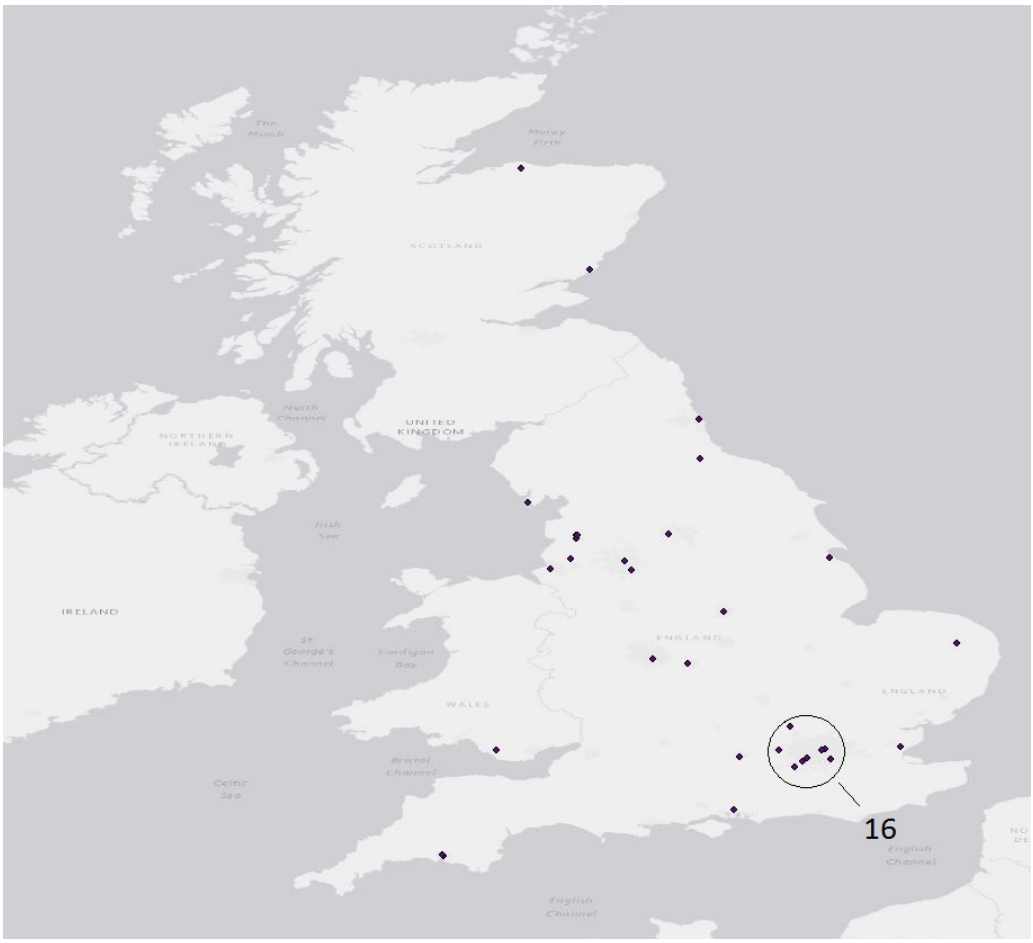

Figure 4. Geolocations of the 42 British Participants in the Strength of Accent Survey

Survey results identified speakers from each country who had similar strength of accent when judged by foreign listeners. One challenge, though, was to select speakers who also had similar strength of accent when judged by listeners from their own country (henceforth "own-accent" rating). This was particularly true of the British speakers, all of whom were scored between 2.9-3.7 on a 7-point scale (7 indicating "very different") by UK listeners; this contrasts sharply with the American and Australian speakers, who scored an own-accent rating of 1.2-2.0 and 1.8-2.6, respectively (see **Error! Reference source not found.**). It is possible that the UK volunteers acquired the US accent more quickly than their Australian counterparts, but this seems unlikely since both Australian speakers had been living in the US for more than two years and three of the UK speakers had been living in the US for little more than six months. However, it is difficult to determine how much—and whether—time in the US affected each speaker without further inquiry. A more likely reason for the high own-accent ratings of these speakers is the possibility that the UK is home to a much broader range of accents and dialects than either the US or Australia, and residents from different parts of the country, therefore, are more likely to sound different to one another.

All three of the speakers who were selected for the study were female. The American speaker was 21 and from Moscow, Idaho. The Australian speaker was 27 and from Sidney, Australia. The British speaker was 18 and from Cheltenham, England. Though Speaker 1 (American) came closer to Speakers 3 (Australian) and 5 (British) in own-accent ratings, Speaker 2 was selected as the American speaker. This was done for two reasons: First, Speakers 2, 3, and 5 were all female, while Speaker 1 was male. Though research suggests that gender of speaker does not influence listening comprehension in TOEFL-style listening tasks (Major *et al.*, 2005), other research demonstrates that a sudden change in speaker can briefly increase the processing

load and response time for the listener (Sommers *et al.* 1994; Clarke & Garrett, 2004). It is reasonable to infer from these findings that a change in speaker gender may possibly affect processing cost more than a simple change in speaker. Second, Speaker 1 regularly used creaky voice during the latter half of his sentences, which made him sound even more distinct from the other speakers. Since EI is dependent on immediate processing and repetition, it was thought best to eliminate variables that might cause unneeded delay or processing on the part of the listener.

Table 1

Accent Rating of Speakers who Volunteered for this Study

Speaker		Mean accent rating by origin of rater				
ID	Origin	US	Australia	UK	Foreign total	Own accent
1	US	2.0	4.6	5.5	5.1	2.0
2	**US**	**1.2**	**4.4**	**5.5**	**5.0**	**1.2**
3	**Australia**	**4.0**	**1.8**	**5.1**	**4.5**	**1.8**
4	Australia	4.3	2.6	4.8	4.6	2.6
5	**UK**	**5.4**	**4.6**	**2.9**	**5.0**	**2.9**
6	UK	4.4	3.1	3.3	3.8	3.3
7	UK	5.2	5.2	3.3	5.2	3.3
8	UK	4.1	4.2	3.7	4.2	3.7
9	UK	6.0	5.6	3.4	5.8	3.4

1 = "Not at all different", 7 = "Very different"

Note: Bolded, underlined rows indicate speakers selected for the study.

It is interesting to note that the American speakers were rated as having lighter accents by the Australian listeners than by the UK listeners. Likewise, the Australian speakers were also rated as having lighter accents by the US listeners. The UK speakers, on the other hand, were rated by both American and Australian listeners as having accents of roughly comparable strength (with the notable exception of Speaker 6). This was surprising since historical ties between the countries suggest that the US would be the odd one out, not the UK. Another interesting point is how low Speaker 2 scored for own-accent rating (1.2, where 1 corresponds to "no accent"). This suggests that perhaps a "standard" accent is more ubiquitous in the US,

though another possible explanation is that the American volunteers had remained in their own country while the other volunteers had not, which may have led to the latter group acquiring foreign accent features. Both of these issues may be fruitful avenues of study for dialectologists.

3.1.2 Test Design. The EI test was composed of two interwoven components: 1) an anchoring (or common) section consisting of 18 items and 2) a section of unique items consisting of 45 items. These items were presented to test-takers in a Latin square design so that the accents would be presented in all orders (see Figure 5); this was to control for the possibility that exposure to one accent would prime the listener for another.

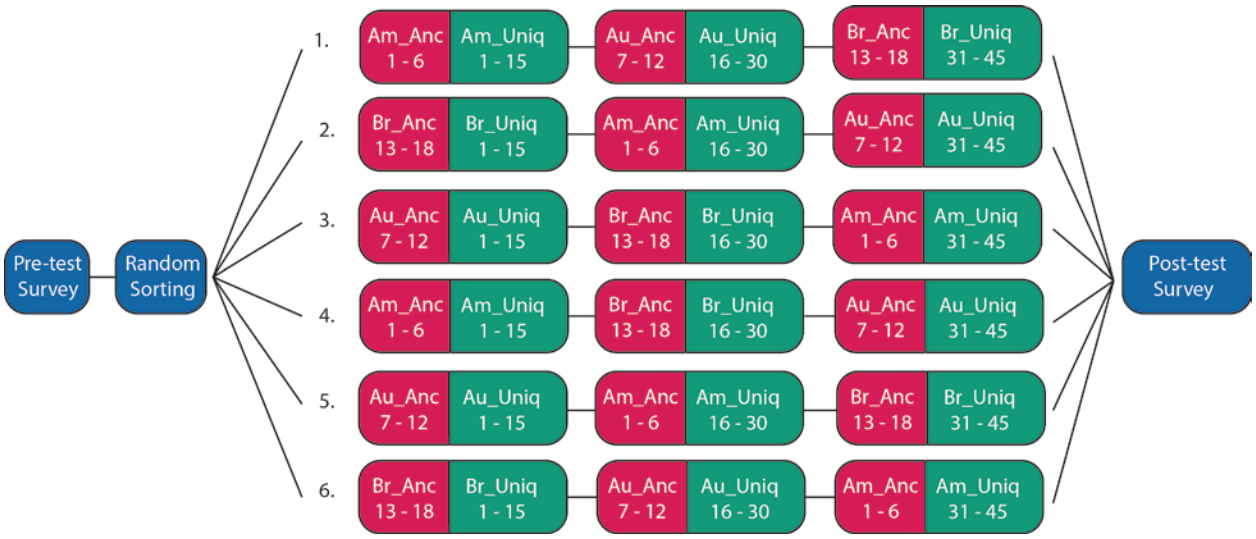

Figure 5. Diagram of EI Test Designs
Note: The first two letters of each file indicate the origin of the speaker (Am = America; Au = Australia; Br = Great Britain). The next set of letters indicates which set of items each file belongs to (Com = common items; Uniq = unique items). The numbers identify each item within each set.

The anchor items were designed so that all participants were exposed to some items in common (all hearing identical recordings). The anchor items were composed of three groups of six prompts, each group being recorded by a speaker of American English, Australian English, or British English. Within each group, the items evenly represented three levels of item difficulty: intermediate, advanced, and superior. These items were created in a similar process as those in

the Cox, Bown & Burdis (2015) study previously cited.

The unique items, on the other hand, were designed to ensure that each group of participants heard an equivalent amount of each accent, but for different items. This was to control for item effect in which the specific features of the EI item might contribute to the score variance more than the accent used. To design these items, each of the three speakers recorded all 45 unique prompts, but test-takers only received 15 prompts from each speaker. Originally, these unique items were intended to be divided evenly between intermediate, advanced, and superior levels for each speaker (as with the anchor items), but an error in the computer programming assigned all intermediate unique items to numbers 1-15, all advanced unique items to numbers 16-30, and all superior unique items to numbers 31-45. The items were still inserted into the various test forms according to the original design, with the result that all unique items for a given accent were of a single difficulty level. This error was not noticed until after data collection and is considered in the final analysis of the data. If the total score had been analyzed using classical test theory, it would have been problematic; however, MFRM calculates item difficulty parameters with person independence. In other words, the difficulty of specific items is calculated probabilistically and the relative location of the item difficulty parameter functions independently of the examinees. When each test form has a set of common items—or anchor items—then it does not matter if some of the items unique to each test form vary in degree of difficulty; the item difficulty parameter can still be computed, and the data is still usable. The implications of this design flaw and how it was adapted to are further discussed chapter 4.

To combine the anchor and unique sections of the assessment, it was organized into three parts according to the accent of each prompt, with anchor items and unique items of a given accent occurring in tandem in single blocks of items. The six anchor items preceded the unique

items in each block. Organizing the items into blocks was for organizational purposes only, and test takers did not experience any pauses between blocks or any other indicator (other than a change of speaker) that the test was organized in such a fashion.

To account for ordering effects of prompt accent, six separate forms of the test were created and administered simultaneously. While the content (i.e. the text) of the prompts was identical for each form, the order in which items appeared was different, as was the accent in which unique items were recorded. For instance, the first part of the assessment consisted of six anchor items followed by unique items 1-15 for all students, but only students assigned to the first and fourth forms heard unique items 1-15 read by an American speaker, and only for these students did anchor items 1-6 (which are always read by an American) precede unique items 1-15; likewise, students in the second and sixth groups also encountered unique items 1-15 in the first block, but these were read by a British speaker and were preceded by anchor items 13-18.

3.3 The Pre-test Survey

Prior to beginning the EI test, participants completed a brief survey about their experience with and exposure to different varieties of English. The survey included five questions, each asking participants to rate their previous exposure to American English, Australian English, British English, other native English varieties, and nonnative varieties of English in different contexts. For each question, participants answered on a 5-point Likert scale for each of the aforementioned accents (see

Table 2). The full survey in the original format of presentation can be found in appendix B.

Table 2

Pre-test Survey Questions

1. Overall, how familiar are you with the following accents?
2. How often do you hear the following English accents on TV, radio, the internet, or other media?
3. How often do you hear the following English accents in face-to-face communication?

4. How long have you studied English with teachers who have the following accents?
5. How long have you lived in the following countries?

3.4 The Post-test Survey

Immediately after finishing the EI test, subjects completed a post-test survey (see Appendix A). This survey included two questions that were designed to investigate the second research question: what are students' perceptions of the effect of accent on EI test scores? The first question explained that the EI test included three speakers, each with a different accent, and asked: "Did any of these accents make it difficult to understand and repeat what you heard?" Students responded to this question on a 5-point Likert scale (1 = "Not at all", 5 = "Very much"). The second question asked students to listen to each of three sound files and to "rate how easy [they] think it is to understand each speaker". The three files were recordings of each speaker reading an identical sentence. Subjects responded to this question with a 5-point Likert scale (1 = "Very easy", 5 = "Very difficult") for each speaker.

3.5 Participants

The study included 232 students at Brigham Young University's English Language Center (ELC), BYU's ESL institution. The student body at the ELC is by design diverse in both language ability, nationality, and L1 background. In terms of English proficiency, students ranged from true beginners to university-ready. Among the participants in this study, 31 countries and 16 languages were represented, with no one country providing more than 15% of the students (see *Table 3*). The diversity of these students provided a good sample for this study; many students had received significant exposure to and instruction in varieties of English other than American before coming to the US.

Table 3

Demographic Information for Research Participants

	Number	Percentage
Gender		
Male	97	42%
Female	135	58%
Age		
18-25	146	63%
26-30	45	19%
31+	36	16%
Place of Origin		
The Americas	156	67%
East Asia	64	28%
Europe	10	4%
Other	2	1%
Native Language		
Spanish	128	55%
Portuguese	27	12%
Korean	19	8%
Japanese	16	7%
Russian	4	2%
Mandarin	4	2%
Other	14	6%

3.6 Test Administration

The test was administered in the ELC's computer lab under the supervision of trained proctors. The computers were arranged in rows and each was equipped with a headset for playback and recording purposes (Sanako model SLH-07).

The EI test began with some sample items and instructions for calibrating the headset, which allowed students to become familiar with the task type and troubleshoot issues with the equipment. Despite this pre-test calibration, however, five students still experienced technical difficulties that made their responses impossible to score. These students were not included in

the final analysis.

Due to the large number of students who completed the task simultaneously—up to 54 at a time—participants were likely to hear a certain degree of background noise in spite of their headsets. Previous research has indicated that adverse listening conditions (e.g. background noise) can interfere with listening comprehension, especially when hearing a novel or unfamiliar accent (Adank *et al.*, 2009). However, since one of the chief advantages of EI testing is the economy of administration, this study was conducted in a typical computer laboratory environment.

3.7 Test Scoring

EI tests have been scored in a variety of ways, including a range of granularity from binary all-correct/incorrect schemes to those that consider each syllable; as for accuracy, some have permitted similar syllables and words while others have required exact repetition (Vinther, 2002). In this study, scores were based on the percentage of syllables correctly repeated without regard for the order of the utterance—this was then converted to a ten-point scale. For example, the hypothetical sentence "I was walking to the store yesterday" includes ten syllables; if a student heard this sentence and then repeated "I walked to the store yesterday" or "I was walking to the store today", the student would score eight out of ten points since two of the ten syllables were missing (in the first case *was* and *-ing*; in the second case *yester-*). Also, the lack of ordering constraints allows students to repair an erroneous repetition: for example, if a student repeated "I went to the store yesterday—I was walking", then full points would be awarded. When compared to other scoring systems, this method allows for a more detailed analysis of test results.

The tests were all scored by trained raters. These were volunteers who were either

graduate students or ELC employees. Training included a detailed explanation of the task that was followed by a supervised rating of fifteen practice items. The scoring was done digitally using a web application developed by the ELC that allowed raters to simultaneously listen to and score each sentence uttered by a student.

3.8 Data Analysis

To answer the first research question regarding the effect of accent on EI test difficulty, the results of the EI test and pre-test survey were analyzed using a Many-Facet Rasch Measurement (MFRM). With MFRM, multiple facets—potential sources of variance—can be accounted for and analyzed to determine whether systematic variance exists for each facet. For example, in this study, accent is included as one facet in order to determine whether accent affected test difficulty in a systematic way. This can be seen by the Rasch separation reliability—a reliability at or near 0 would indicate that variance is not systematic, and a reliability of at or near 1 would indicate that it is. In this study, the facets analyzed were (1) examinees, (2) accent, (3) test items, and (4) test form (order of accent). Since the focus of this study was whether accent systematically affected test difficulty, a crucial part of the reported data is the separation reliability for the accent facet.

It was also expected that the test form facet would have a separation reliability of near 0. This would indicate that test form had no systematic effect on test difficulty, which was the aim of the test design. Both the examinee and test item facets were expected to have a separation reliability of near 1 as it is expected that examinees vary in ability and test items vary in difficulty.

To perform the analysis, the FACETS software package was used (Linacre, 2011). This program also produces a display of data known as a *vertical scale* (also known as a *variable*

map). Vertical scales depict each facet on a single, comparable scale. The measure to which each facet is calibrated is the *logit* (log odds ratio). With logits, different—but interrelated—concepts such as student ability or test item difficulty can be compared. One benefit of MFRM is that when systematic variance is present and that variance can be accounted for by using the logit measure or the Fair Average. Thus, if accent were to systematically affect the scores examinees, then MFRM could model out the variance attributable to that facet. If it were not to affect the outcome systematically, then future test designers would not need to account for accent as a facet that affected the outcome.

If accent proved to be a factor in EI test item difficulty, it would be important to better understand how this is affected by accent familiarity. In the case that accent accounted for systematic variance in test scores, an ANOVA would be used between the results of test scores of items of a given accent and the pre-test survey question "Overall, how familiar are you with the following English accents?".

To answer the second research question regarding how students perceive the effect of accent on EI test difficulty, the post-test survey results were analyzed to determine how the body of students felt about accented items overall. To answer the second part of this question—how accurate student perceptions were—an ANOVA was done between EI test scores and the self-reported degrees of accent familiarity from the pre-test survey.

4. Results

This chapter provides an analysis of the data gathered in this study. The research questions are discussed separately and in order.

4.1 Research Question 1

The first research question, "What effect does accent have on EI test difficulty?" was addressed by using MFRM to analyze the test results and the pre-test survey results. Some participants experienced technical or user-related difficulties with their headsets and their test responses were not scorable. For some participants, this only affected a few items, but for others it affected all of their items. The responses of 5 out of 232 original participants were completely unscorable, leaving the scores of 227 participants for the final analysis. An exact breakdown of how many participants took each form of the test can be seen in *Table 4*.

Table 4

Number of Participants per Test Form

Test Form	Number of Participants
1	36
2	35
3	41
4	40
5	38
6	37
Total	227

With a Many Facets Rasch Measurement analysis, facets are compared on a vertical scale (see Figure 6). The facets considered in this analysis were (1) examinee, (2) accent, (3) item, and (4) test form. The first column on the vertical scale indicates the logit measurement, which is based on the mean performance of the examinees (the mean is indicated by 0).

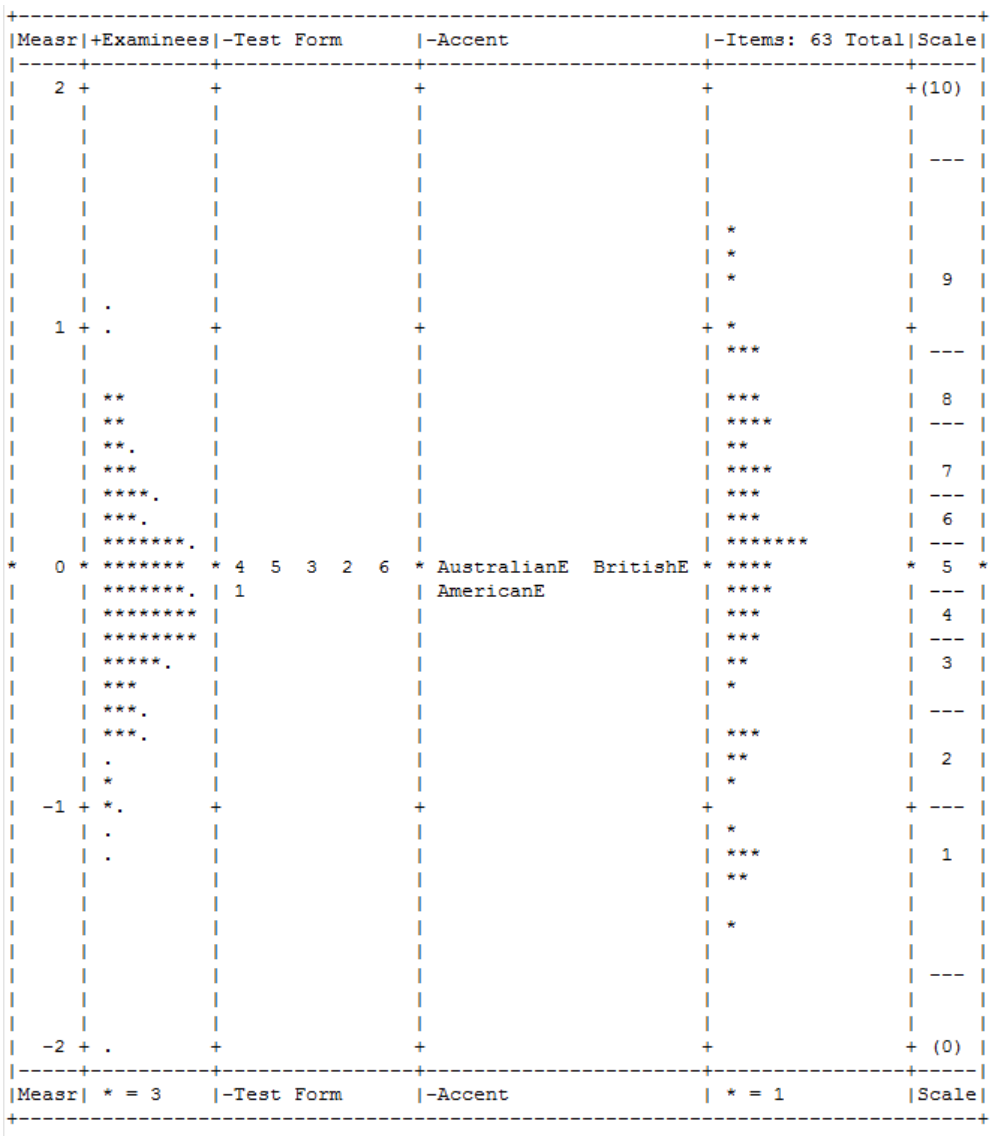

Figure 6. Vertical Scale of the Results of Many Facets Rasch Measurement.
Note: Columns: (1) logits, (2) examinee performance, (3) test form (order of accent), (4) accent, (5) test item, (6) scale

As can be seen in the vertical scale, the results indicated that accent did have an effect. Specifically, Australian and British accents negatively impacted the scores of most participants. A closer examination of the data shows that the logit for American English was -0.06, Australian English was 0.01, and British English was 0.05 (see **Error! Reference source not found.**) and this facet appeared to function predictably as all of the Outfit values fell within the expected range of .5 and 1.5. Since MFRM models out systematic variance from different facets, the

person ability estimate factors out the effect accent has on the person logit. If classical test theory were used, however, it would mean that if a given student took this test with prompts in American English and scored a 48% (the Observed Average from *Table* 5 converted back to percent syllables correct), that student would likely score 46% on the same test if the items were recorded in an Australian accent and 43% if the items were recorded in a British accent. For this set of test items, accent could account for up to 5% variation in test score. In addition, the accent facet had a separation reliability of .98, indicating that this variation in test difficulty was systematic.

Table *5*

MFRM Measurement Report for Accent

Accent	Syllables Correct	Observed Average Scale	Total Count	Logit	Outfit Mean Square	Fair Average
American English	48%	4.83	4673	-0.06	1.03	4.63
Australian English	46%	4.61	4646	0.01	1.00	4.33
British English	43%	4.34	4659	0.05	1.01	4.13
Mean	45%	4.59	4659.3	0.00	1.01	4.36
SD	2%	.20	11	.04	.01	.20

The test form facet also had a high separation reliability (.90), meaning that test form affected overall test difficulty. This is generally avoided in test design and is likely due to the error encountered in designing this test. That is, each of the forms had all of the intermediate items in the same accent (e.g. American English) instead of having the intermediate items equally distributed among all three accents, and the same issue happening with the Advanced and Superior items. In examining the test forms more closely (see Table 6), the easier forms were those in which the Intermediate items were spoken in an American English Accent (Forms 1 and 4), and thus it was more likely the students would receive higher scores on those test forms. Since the Advanced and Superior items were likely too difficult for many examinees to answer

correctly regardless of the accent, the advantage of the American accent was negated. Thus, if this analysis had been conducted with classical test theory, this design flaw could have had a greater impact on the analysis. However, since MFRM is person and item independent, this systematic variance can be controlled for

Table 6

MFRM Measurement Report for Test Form (Accent Order) Ordered by Logit

Test Form (Accent Order)	Syllables Correct	Observed Average Scale	Total Count	Logit	Outfit Mean Square	Fair Average
1 (AmAuUK)	49%	4.92	2199	-0.05	1.07	4.62
4 (AmUKAu)	46%	4.59	2432	-0.03	1.07	4.49
6 (UKAuAm)	47%	4.66	2263	0	1.02	4.38
5 (AuAmUK)	46%	4.58	2344	0.01	1.06	4.31
2 (UKAmAu)	44%	4.41	2169	0.04	0.98	4.2
3 (AuUKAm)	44%	4.43	2571	0.04	0.9	4.18
Mean	46%	4.36	2329.7	0	1.02	4.36
SD	2%	0.16	139.2	0.03	0.06	0.16

Table 7

Separation Reliability Statistics for Examinees, Accent, Items, and Test Form

	Examinees	Accent	Items	Test Form
	N = 227	N = 3	N = 63	N = 6
Measures				
Mean	.77	.77	.51	.79
SD	.09	.02	.11	.05
Outfit				
Mean	1.04	1.01	1.01	1.02
SD	0.55	0.02	0.27	0.07
Separation statistics				
Separation Reliability	.94	.98	1.00	.90
Strata Index	5.68	8.92	23.04	4.32

4.2 Research Question 2

The second research question, “What are students’ perceptions of the effect of accent on

EI test difficulty?" was addressed by analyzing the post-survey data; to explore the related question, "how accurate are these perceptions?" an ANOVA was conducted between the EI test scores and students' self-reported accent familiarity from the pre-test survey.

A glitch in the test design allowed some students to bypass both the pre-test and post-test surveys, so survey results were not obtained for all of the 232 participants. Only 146 pre-test surveys and 178 post-test surveys were completed.

The post-test survey included two questions. One question asked participants whether any of the accents they heard made it difficult to understand and repeat what they heard; the second question asked participants to listen to three audio samples of different speakers and rate how easy each was to understand. These three samples were recordings of the speakers from the survey reading an identical sentence (the sentence was not included in the EI test).

An analysis of these survey results indicates that students perceive that extra difficulty is introduced by accented speech. In response to the first question "Did any of these accents make it difficult to understand and repeat what you heard?", students largely thought that they did. On a scale of 1 ("Not at all") to 5 ("Very much"), the overall mean was 3.89 (N = 178, SD = 1.16, 95%CI [3.72, 4.06].

Responses to the second question, "Rate how easy you think it is to understand each speaker," also suggested that these test takers thought that the British speaker was more difficult to understand than the Australian speaker, who was in turn more difficult to understand than the American speaker. On a scale of 1 ("Very Easy") to 7 ("Very Difficult"), the mean difficulty rating was 3.05 for the British speaker, 2.51 for the Australian speaker, and 2.02 for the American speaker (see *Table 8*).

Table 8

Ease of Understanding Accent Descriptive Statistics

	American English	Australian English	British English
Mean	2.02	2.51	3.05
SD	1.56	1.57	1.85
95% CI	[1.79, 2.25]	[2.28, 2.74]	[2.78, 3.32]

Note: N = 178

When students were divided into bands of familiarity based on the pre-test survey, it was found that most students were familiar with American English but less familiar with Australian and British English. Few students claimed to be familiar with Australian or British English (see Figure 7).

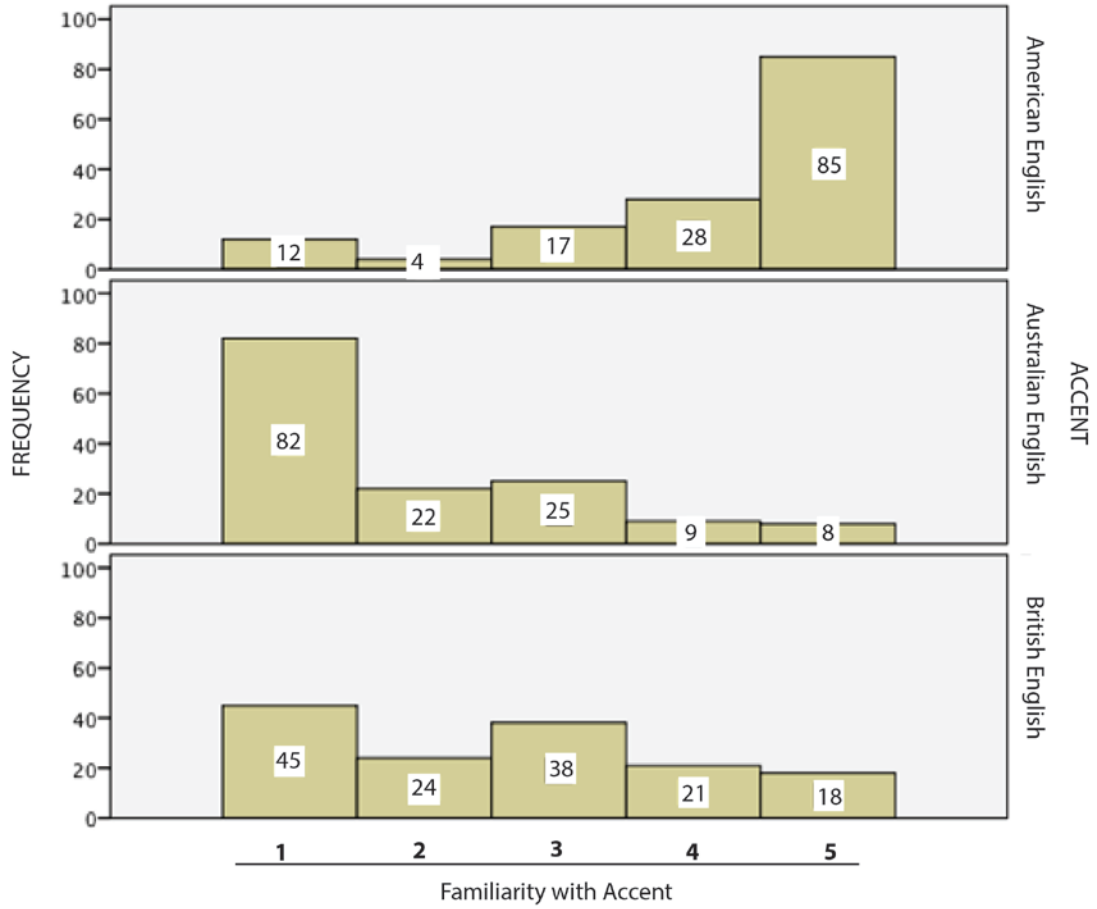

Figure 7. Results of Pre-Test Survey
Note: Responses are to the question "Overall, how familiar are you with the following English accents?". 1 = "Not at all Familiar"; 5 = "Familiar".

An examination of the EI observed average test scores found that the more familiar examinees were with a particular accent, the more likely they were to have a higher test score (see *Table 9*). This was particularly true for those that were Familiar (Likert Scale category 5) compared to those that were Not at all Familiar (Category1).

Table 9

Descriptive Statistics of EI Observed Average Test Score by Accent Familiarity

ACCENT		N	Mean	SD	95%CI
American English	1—Not at all Familiar	12	3.62	1.71	[2.53, 4.71]
	2	4	4.71	1.84	[1.78, 7.65]
	3	17	3.97	1.64	[3.12, 4.82]
	4	28	4.75	1.73	[4.08, 5.42]
	5—Familiar	85	5.38	2.05	[4.94, 5.82]
	Total	146	4.93	1.99	[4.61, 5.26]
Australian English	1—Not at all Familiar	82	4.35	1.95	[3.92, 4.77]
	2	22	5.24	1.70	[4.48, 5.99]
	3	25	5.07	2.01	[4.25, 5.90]
	4	9	5.17	1.47	[4.04, 6.30]
	5—Familiar	8	5.80	2.14	[4.02, 7.59]
	Total	146	4.74	1.94	[4.42, 5.05]
British English	1—Not at all Familiar	45	3.95	1.54	[3.48, 4.41]
	2	24	4.26	2.02	[3.41, 5.11]
	3	38	4.55	2.00	[3.89, 5.21]
	4	21	5.30	1.69	[4.53, 6.07]
	5—Familiar	18	5.43	2.11	[4.38, 6.48]
	Total	146	4.53	1.90	[4.22, 4.84]

To determine if this finding was statistically significant, three one-way ANOVAs were conducted between accent familiarity (independent variable) and the EI observed average test scores (dependent variable). Familiarity with American English was a significant factor in the EI test scores, ($F(4, 145) = 13.68$, $p = .007$), though the only significant difference was between

those with low and medium familiarity (1 and 3 on a scale of 1-5) and those with high familiarity (5 on a scale of 1-5). The LSD mean difference between familiarity levels 1 and 5 was -1.76, 95% CI [-2.93, -0.592], p = .003, Cohen's *d* = .93, thus the effect size between being *Not at all Familiar* (1) and *Familiar* (5) was large. The LSD mean difference between familiarity levels 3 and 5 was -1.41, 95% CI [-2.42, -0.40], p = .006, Cohen's *d* = .76, thus the effect size between being somewhat *Familiar* (3) and *Not at all Familiar* (5) was medium to large. It is interesting to note that there was not a significant difference between categories 2 and 5. This is probably an artifact of only having 4 participants that had selected that category.

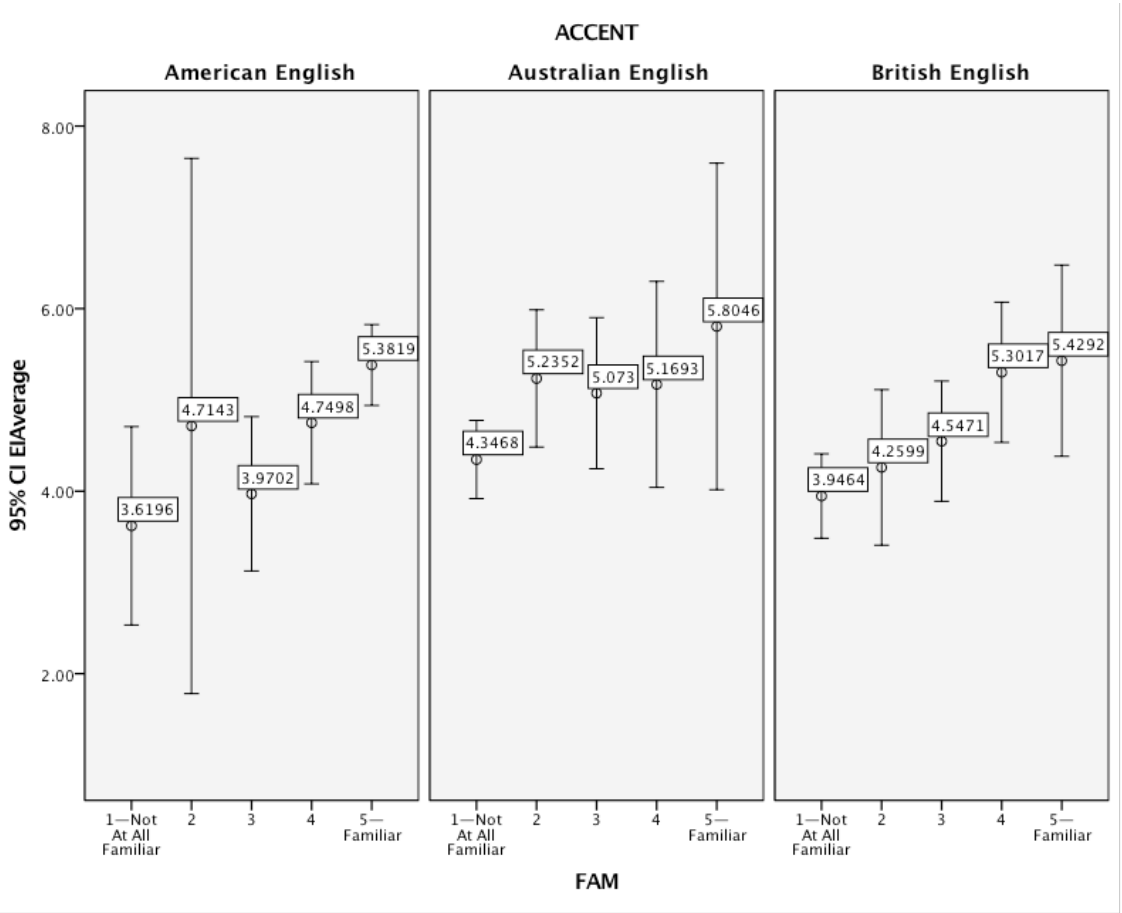

Figure 8. Mean Scores with 95% Confidence Intervals of EI Test Accented Portions Based on Self-Reported Familiarity of Accent

Familiarity with British English was also a significant factor in test scores (F (4, 145) = 11.04, p = 0.014), though again, the only significant difference was between those with low

familiarity (1 and 2 on a scale of 1-5) and those with high familiarity (5 on a scale of 1-5). The LSD mean difference between the extremes (categories 1 and 5) was -1.48, 95% CI [-2.49, -0.47], p = .005, Cohen's *d* = .80, thus the effect size between being *Not at all Familiar* (1) and *Familiar* (5) was large. The LSD mean difference between the categories 2 and 5 was -1.16, 95%CI [-2.30, -0.03], p = .04, Cohen's *d* = .56, thus the effect size between being not very *Familiar* (2) and *Familiar* (5) was medium.

Familiarity with Australian English, however, was not a significant factor in the EI test scores overall (F (4, 145) = 7.89, p = .076). The middle three categories (2, 3 and 4 on a scale of 5) had very similar means, but the extreme categories of low familiarity (1 on a scale of 1-5) and high familiarity (5 on a scale of 1-5) were significantly different with a LSD mean difference = -1.45, 95% CI [-2.86, -.06], p = .04, Cohen's d = .71, thus the effect size between being *Not at all Familiar* (1) and *Familiar* (5) was medium

5. Discussion

5.1 Review of Findings

The first research question investigated the effect of accent on EI test difficulty. The results found that accent does have an effect on item difficulty. For the group of students in this study, British English was the most difficult, Australian English was the second most difficult, and American English was the easiest.

The second research question investigated students' perceptions of whether accent affected test difficulty and how accurate these perceptions were. The results indicated that most students thought that accent had an effect; specifically, the students in this study felt that British English was harder than Australian English, which was in turn harder than American English. An ANOVA that was done between the test scores and self-reported accent familiarity indicated that overall, perceptions of difficulty aligned with actual difficulty: students who claimed to be unfamiliar with an accent tended to perform more poorly on items in that accent.

5.2 Implications

Since it has been demonstrated that accent can have an effect on EI test results, designers of this kind of language assessment must consider accent when creating audio prompts. This is of particular importance since EI testing is gaining ground as a low-cost alternative or supplement to some traditional forms of low-stakes assessment. For world languages such as English or Spanish, language programs must consider the wide variety of accents that their students may have been primarily taught, which, as this study finds, may place some students at a disadvantage if accent is not accounted for in EI testing. If an institution opted for single-accent EI tests, it may be beneficial to recognize and provide a justification for the increase in difficulty to students unfamiliar with that accent. If an EI test is administered at an institution (as opposed to online),

for example, it could be argued that the test is designed to assess language ability in a local context. On the other hand, an assessment that is designed to be administered world-wide to a global audience may benefit from using a variety of accents. This would ensure that the increase in difficulty brought on by novel accents would be shared by all test-takers. Whatever approach is taken, test administrators may help test-takers by providing an explanation and justification for the accents used in EI recordings. This may help mitigate a decrease in motivation that can occur when a test has low face validity.

This study also has implications for listening assessment generally. Previous studies have already demonstrated that accent can affect listening comprehension in a TOEFL setting (e.g. Ockey & French, 2014)—that is, with tasks that require participants to listen to passages and answer multiple-choice questions about the content—but the results of the current study indicate that accent affects comprehension in a different type of listening assessment. It is possible that accent affects difficulty in other types of listening assessment tasks.

Though the findings of this study confirm previous research that suggests that accent can impair listening comprehension (e.g. Varonis & Gass, 1982; Anderson-Hsieh & Kohler, 1988; Floccia *et al.*, 2006), it also adds to the very limited body of research on the effect of regional or international accent in listening in a second language and confirms the findings of the previous studies (Ockey & French, 2014; Major *et al.*, 2005). Furthermore, it is the first study to explore accent and listening in the context of EI testing, which is a unique way of measuring listening comprehension. Unlike other studies that measure the effect of accent on response time or judgment about content, elicited imitation measures participants' ability to repeat a sentence—a process that can be done without a full understanding of all lexical items. The results of this study therefore suggest that accent may affect more than the speed or accuracy at which a

judgment is made on the part of the listener.

5.3 Limitations

There are some limitations to this study that should be considered. First, Derwing and Munro's (2009) definition of accent views it as a feature of the observer's perception rather that a feature of speech. In this study, native speakers were used to determine strength of accent, and it was assumed that this represented how the speech samples were perceived by nonnative speakers from a wide range of proficiency and L1 backgrounds. No studies have yet confirmed that listeners perceive an L2 accent in the same way that they perceive an L1 accent.

Second, rate of speech was not accounted for in the creation of the EI prompts. Previous research has demonstrated that rate of speech is a factor that can affect listening comprehension (Anderson-Hsieh & Kohler, 1988). It was not controlled for in the present study because it would have required either a good deal of training and practice on the part of the volunteers or a manipulation of the audio files; the cost of the former was beyond the scope of this study, and the latter had the potential risk of introducing unnatural distortions into the audio.

Finally, it is possible that the accent of the audio prompt had an effect on how some test-takers pronounced the repetition and, therefore, how their responses were scored. During the scoring process, raters noted that some words were pronounced in a distinctly non-American way; it is impossible to say whether this was part of the speakers' normal speech patterns or the result of a particularly good imitation of the British or Australian speaker, but perhaps future research may provide illumination on the topic.

5.3 Future Research

Though not central to the research questions of this study, it was found that for the Australian listeners who participated in the preliminary accent survey our American speakers

were more similar to Australians in terms of accent than the British speakers. Likewise, the Australian volunteers were more similar to American listeners than were British speakers. These findings may provide an interesting point of departure for related studies in perceptual dialectology.

5.4 Conclusion

This study examined (1) the effect of accent on elicited imitation (EI) test difficulty and (2) student perceptions of this effect. A Many Facets Rasch Measurement was used to analyze the test results, and it was found that different accents did affect the difficulty of test items. It was also found that participants—who were all familiar with American English—perceived that Australian and British accented prompts made the elicited imitation task more difficult.

References

Abeywickrama, P. (2013). Why Not Non-native Varieties of English as Listening Comprehension Test Input?. *RELC Journal*, *44*(1), 59-74.

Adank, P., & McQueen, J. M. (2007). The effect of an unfamiliar regional accent on spoken word comprehension.

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 520.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language learning*, *42*(4), 529-555.

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language learning*, *38*(4), 561-613.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford University Press.

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology*, *82*(2), 300.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647-3658.

Cox, T. L., Bown, J., & Burdis, J. (2015). Exploring Proficiency-Based vs. Performance-Based Items With Elicited Imitation Assessment. *Foreign Language Annals*, *48*(3), 350-371.

Cox, T., & Davies, R. S. (2012). Using automatic speech recognition technology with elicited oral response testing. *Calico Journal*, *29*(4), 601-618.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility. *Studies in second language acquisition*, *19*(01), 1-16.

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*(04), 476-490.

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing?. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1276.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language learning*, *34*(1), 65-87.

Graham, C. R., Lonsdale, D., Kennington, C. R., Johnson, A., & McGhee, J. (2008, May). Elicited Imitation as an Oral Proficiency Measure with ASR Scoring. In *LREC*.

Linacre, J. (2011). Facets many-facet Rasch measurement software.

Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL quarterly*, *36*(2), 173-190.

Major, R. C., Fitzmaurice, S. M., Bunta, F., & Balasubramanian, C. (2005). Testing the effects of regional, ethnic, and international dialects of English on listening comprehension. *Language learning*, *55*(1), 37-69.

Matsushita, H., & Lonsdale, D. (2012). Item Development and Scoring for Japanese Oral Proficiency Testing. In *LREC* (pp. 2682-2689).

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543-562.

Millard, B., & Lonsdale, D. (2014, December). French oral proficiency assessment. In *Variation*

within and across Romance Languages: Selected papers from the 41st Linguistic Symposium on Romance Languages (LSRL), Ottawa, 5–7 May 2011 (Vol. 333, p. 401). John Benjamins Publishing Company.

Moulton, Sara E., "Elicited Imitation Testing as a Measure of Oral Language Proficiency at the Missionary Training Center" (2012). *All Theses and Dissertations.* Paper 3137. http://scholarsarchive.byu.edu/etd/3137

Mugglestone, L. (2007). *Talking proper: The rise of accent as social symbol*. Oxford University Press, USA.

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, *45*(1), 73-97.

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and speech*, *38*(3), 289-306.

Ockey, G. J., & French, R. (2014). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*, amu060.

Okura, E., & Lonsdale, D. (2012). Working memory's meager involvement in sentence repetition tests. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 2132-2137).

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *The Journal of the Acoustical Society of America*, *96*(3), 1314-1324.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, 0265532211424478.

Varonis, E. M., & Gass, S. (1982). The comprehensibility of non-native speech. *Studies in*

second language acquisition, *4*(02), 114-136.

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*(1), 54-73.

Wingstedt, M., & Schulman, R. (1984). Comprehension of foreign accents. *Phonologica*, 339-345.

Appendix A

Pre-test Survey

Overall, how familiar are you with the following English accents?

	Not at all familiar				Familiar
American (US)	○	○	○	○	○
Australian	○	○	○	○	○
British (UK)	○	○	○	○	○
Other native accents (eg. Canadian, New Zealander, etc.)	○	○	○	○	○
Non-native accents	○	○	○	○	○

How often do you hear the following English accents on TV, radio, the internet, or other media?

	Rarely				Very often
American (US)	○	○	○	○	○
Australian	○	○	○	○	○
British (UK)	○	○	○	○	○
Other native accents (eg. Canadian, New Zealander, etc.)	○	○	○	○	○
Non-native accents	○	○	○	○	○

How often do you hear the following English accents in face-to-face communication?

	Rarely				Very often
American (US)	○	○	○	○	○
Australian	○	○	○	○	○
British (UK)	○	○	○	○	○
Other native accents (eg. Canadian, New Zealander, etc.)	○	○	○	○	○
Non-native accents	○	○	○	○	○

How long have you studied English with teachers who have the following accents?

	Not at all	Less than 1 year	1-2 years	3-4 years	5+ years
American (US)	○	○	○	○	○
Australian	○	○	○	○	○
British (UK)	○	○	○	○	○
Other native accents (eg. Canadian, New Zealander, etc.)	○	○	○	○	○
Non-native accents	○	○	○	○	○

How long have you lived in the following English-speaking countries?

	Not at all	Less than 6 monts	6-12 months	1-2 years	3+ years
United States	○	○	○	○	○
Canada	○	○	○	○	○
Australia	○	○	○	○	○
United Kingdom	○	○	○	○	○
Other (please be specific)	○	○	○	○	○

Post-test Survey

In the test you just completed, you listened to three speakers, each with a different accent.

Did any of the accents you listened to negatively affect your ability to understand and repeat what you heard?

	Not at all				Very much
	○	○	○	○	○

Listen to each of the audio files below.

Rate how easy you think it is to understand each speaker.

	Very easy						Very difficult
	○	○	○	○	○	○	○
	○	○	○	○	○	○	○
	○	○	○	○	○	○	○

May we contact you later for further discussion regarding this survey and the test?

If you would like to be contacted, please select 'Yes' below and provide your name in the blank space.

You are not required to select 'Yes'. Selecting 'Yes' and providing your name will waive your right to confidentiality.

○ Yes

○ No