2014-03-17

Designing and Evaluating a Russian Elicited Imitation Test to Be Used at the Missionary Training Center

Jacob R. Burdis
Brigham Young University - Provo

Designing and Evaluating a Russian Elicited Imitation Test to Be Used at the Missionary Training Center

Jacob R. Burdis

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

Doctor of Philosophy

Richard E. West, Chair
Peter Rich
Troy Cox
Jennifer Bown
Ray Graham

Department of Instructional Psychology and Technology

Brigham Young University

March 2014

ABSTRACT

Designing and Evaluating a Russian Elicited Imitation Test to be Used at the Missionary Training Center

Jacob R. Burdis
Department of Instructional Psychology and Technology, BYU
Doctor of Philosophy

Elicited Imitation (EI) is an assessment approach that uses sentence imitation tasks to gauge the oral proficiency level of test takers. EI tests have been created for several of the world's languages, including English, Spanish, Japanese, French, and Mandarin. Little research has been conducted for using the EI approach with learners of Russian. This dissertation describes a multi-faceted study that was presented in two journal articles for the creation and analysis of a Russian EI test. The EI test was created for and tested with Russian-speaking missionaries and employees at the Missionary Training Center (MTC) in Provo, UT. The first article describes the creation of the test and analyzes its ability to predict oral language proficiency by comparing individuals' scores on the EI to their scores on the Oral Proficiency Interview (OPI). The test was found to effectively predict an individual's OPI score ($R^2 = .86$). The second article analyzes the difference in person ability estimates and item difficulty measures between items from a general content bank and a religious content bank. The mean score for the content specific items ($\bar{x} = .51$) was significantly higher than the mean score for the general test ($\bar{x} = .44$, $p < 0.001$). Additionally, the item difficulties for the religious items were significantly less than the item difficulties for the general items ($p < 0.05$).

Keywords: elicited imitation, elicited response, oral proficiency assessment, language assessment, language for specific purposes, Russian language assessment

ACKNOWLEDGEMENTS

This project was made possible by the efforts of the administration of the Missionary Training Center, the faculty of the Instructional Psychology and Technology Department, the Center for Language Studies, and the Germanic and Slavic Department at Brigham Young University. Ken Packer and Ray Graham were instrumental in providing the feedback and support needed to accomplish this study. Richard West was immensely helpful in assisting in the collaborating and writing of the research study. I also express sincere gratitude to Troy Cox for his patience and guidance throughout the course of this project. Jennifer Bown and Peter Rich were also very helpful in providing support and encouragement needed to finish this project.

Last but not least, I express gratitude for my wife Christie, who sacrificed time, energy, and often sleep to help support me during the course of this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

Elicited Imitation (EI) is a language assessment approach in which test-takers are required to listen to sentences in the target language and without delay repeat back what they hear. The premise of EI is that learners must comprehend what was heard and reproduce it using the language skills available to them. As sentences increase in complexity, learners are unable to simply parrot the sounds of the sentence. They must be able to decode the sentence for meaning and recreate the meaning using language productive skills. Thus, the more comprehension and productive skills available to the learner, the better he or she is able to reproduce what was heard.

EI boasts many benefits over other forms of language assessment. First, administering an EI test is very quick. A typical EI test can be administered in under 15 minutes (depending on the number of items). We are unaware of another language proficiency test that can be administered in such a short amount of time and remain both valid and reliable. The Oral Proficiency Interview (OPI) is the golden standard for assessing oral language proficiency, and it takes 30 to 45 minutes to administer, not including the scheduling burden needed to set up the telephone interview. Next, scoring an EI test is also very quick. A human rater of an EI test can typically score an EI test in less than half the time it took to administer. All that is required for a human rater in scoring an EI test is to identify whether each unit of measure (typically a syllable) in the item was said correctly or incorrectly. Several EI tests have begun to use automated speech recognition (ASR) software to automatically score EI tests. This cuts the already small amount of time needed to score an EI test to a fraction. The OPI requires a highly specialized, skilled rater to assess a test-taker's language proficiency.

While the residual benefits of efficient administration and scoring make EI an attractive assessment alternative, the upfront burden of creating an effective EI test scare many away. EI

tests must be carefully crafted and evaluated before they can be deemed both valid and reliable. Much care must be taken to ensure that the items used in the EI test actually differentiate between test-takers of various proficiency levels. Factors such as sentence length, grammatical complexity, and lexical difficulty must be considered when choosing which items to include.

This study consists of two journal articles that are designed to discuss the factors that need to be considered when constructing EI test items. The goal of the first article is to implement procedures that have been successful in the literature in creating EI test items and apply them to a language that has received little attention in regards to EI—Russian. After creating the Russian EI test, we had test takers take the Russian EI test and a Russian OPI. We compared the scores of these tests in order to discover EI's ability to predict the test-takers' oral language proficiency as made evident by their scores on the OPI. The goal of the second article is to investigate in more depth the effect of content in the items on the person's ability estimates and item difficulty scores of an EI test. We created an EI test in which half the items came from a general content domain, and half came from a content domain familiar to those taking the test.

Both of these articles stem from the same study. We created the EI test to be used with Russian-speaking missionaries and employees of the Missionary Training Center (MTC) in Provo, UT. The test consisted of 72 items, 36 of which originated randomly from a corpus of spoken Russian, and 36 of which came from a corpus of religious stories and testimonies collected from the website Mormon.org/rus. There were ninety-six participants in the study: 52 missionaries studying Russian in the MTC and 44 MTC employees who had recently returned from their missionary service in a Russian-speaking country. Of the MTC employees, three were native Russian speakers. The Russian EI test is the third test (following English and Spanish) developed and tested at the MTC. Because administering and scoring EI is so efficient, and the

volume of language learners at the MTC is so great, the MTC is pursuing creating EI tests in at least 10 of its major languages. The procedures learned from this study will provide helpful guidance to the MTC and to the field as a whole in the creation of EI tests for other languages.

Description of Structure and Content

The format of this dissertation follows the hybrid dissertation model supported in Brigham Young University's McKay School of Education. This model differs from the traditional "five chapter" dissertation approach in that it focuses on producing two journal-ready manuscripts. As a result, the format of this dissertation presents the journal-ready manuscripts as the centerpiece. Each manuscript is its own chapter, and following the manuscripts are the manuscript's reference section and the appendices. Chapter one provides and introduction to the research conducted in these manuscripts as a whole. Chapters two and three present the journal-ready manuscripts. Chapter four provides a conclusion to tie together the research presented in each of the manuscripts. This dissertation also presents a reference section that represents the references for both of the manuscripts.

The target journal for the first article is *The Modern Language Journal*. The mission of the journal is to publish research and discussion about the teaching and learning of foreign languages. The journal follows the APA style format and the manuscript length for submission is 8,000 to 10,000 words. The target journal for the second article is *Language Testing*. The mission of this journal is to publish research and discussion on the fields of first and second language testing and assessment. The journal follows the APA style format and the manuscript length is 4,000 to 8,000 words.

Chapter 2: Article One

Elicited Imitation as a Predictor of Language Proficiency for Learners of Russian

Jacob Burdis

Richard E. West

Troy Cox

Jennifer Bown

Brigham Young University

ABSTRACT

This study investigates the creation and use of a Russian language assessment that predicts oral language proficiency as measured by the Oral Proficiency Interview (OPI). Elicited imitation is a language assessment method that requires test-takers to repeat sentences in the target language. The accuracy at which test-takers are able to repeat more difficult sentences indicates the test-takers' language proficiency. This paper documents the study of 54 students of an intensive 9-week Russian language learning program and 44 learners of Russian who have recently returned from extensive experiences abroad in Russian-speaking countries. This project furthers the research by investigating the validity of an EI test in a new language—Russian. This study found that the scores on an EI test could effectively predict the test-taker's OPI score ($R^2 = 0.86$), giving further evidence of the validity of the EI assessment approach.

Elicited Imitation as a Predictor of Language Proficiency for Learners of Russian

Assessing oral proficiency in a foreign language has traditionally been a difficult and time-intensive task. In fact, many foreign language tests focus on measuring skills that are easier to measure (e.g., knowledge of grammar rules and patterns) in order to avoid the cost and time it takes to measure oral language proficiency. The traditional method of measuring oral language proficiency is conducting a language-speaking interview. This is a very time-consuming and labor-intensive approach. In order to get an accurate measurement, the speech samples taken from the language learner during the interview are recorded and then rated by at least two qualified raters using a well-constructed rubric. Most language learning institutions do not have the time or resources to engage in this process with any degree of regularity.

Commercial options are available to measure oral language proficiency at a high cost. For example, Language Testing International administers an oral proficiency interview (OPI), which is regarded among the language learning community as a valid and reliable measure of oral language proficiency (Radloff, 1991). The OPI consists of a 30 to 45-minute interview conducted over the phone and professional evaluation of the recorded speech sample. The cost of a single OPI is approximately 130 U.S. dollars, making the OPI not feasible for regular oral language proficiency assessment for most language learning institutions. There is clearly a need for a simpler, more affordable approach to measuring oral language proficiency.

In an attempt to meet this need for learners of Russian, this study created a valid instrument utilizing the elicited imitation (EI) approach and investigate its ability to accurately and reliably predict the oral language proficiency of Russian language learners as made evident by their scores on the Oral Proficiency Interview (OPI). EI is an assessment approach in which test takers listen to items in the target language, and then repeat back exactly what they hear.

The accuracy at which they repeat the sentence indicates the test-taker's oral language proficiency. EI instruments are much less expensive to administer than traditional proficiency tests. By implementing effective EI instruments, language-learning institutions could greatly decrease the cost of language proficiency evaluations. To date, very little has been published in the literature regarding the creation of an EI instrument for learners of Russian. Establishing EI as a valid assessment that effectively predicts language proficiency has large implications for the language learning community.

Literature Review

Elicited imitation varies significantly from traditional methods in how it assesses language ability. Traditional language assessment measures focus on creating questions that gauge learners' grammatical competence, vocabulary breadth and depth, and performance on the four language skills (speaking, listening, reading, and writing). They typically use a variety of test items, including but not limited to multiple choice, fill in the blank, matching, error identification and correction, short answer, essay, and verbal response (Bernstein, Van Moere, & Cheng, 2010). EI is an assessment tool that is designed to measure language ability by utilizing a single item type—sentence repetition. Learners are required to listen to a sentence and repeat back exactly what they hear (Chaudron, Prior, & Kozok, 2005). They continue this process as they are confronted with increasingly longer sentences. They are graded by the accuracy of their imitations. It is important to note that EI is a criterion-referenced test rather than a norm-referenced test, as performance is measured against a standard of accuracy that is independent of reference to the performance of others (Brown & Hudson, 2002). EI is a highly intriguing approach because of the relatively small number of resources needed to facilitate it—in several languages it has been done through a computer and speech recognition technology (Cook,

McGhee, & Lonsdale, 2011; Graham, Lonsdale, Kennington, Johnson, & McGhee, 2008). For a more complete review of EI and its history in the literature, consult Vinther (2002) and Bley-Vroman & Chaurdon (1994).

What Does EI Measure?

The purpose of this study is to investigate whether the EI test can predict the complexity of language that exists in a language proficiency scale. More specifically, to determine whether items based on an established proficiency scale will predict test-takers' score on a language proficiency assessment such as the OPI. To begin this investigation, it is important first of all to understand what EI actually measures. Many proponents of EI tests have made claims that EI tests are reconstructive in nature and provide a representation of a learner's interlanguage system. Although EI doesn't directly measure oral language proficiency, it can be used to predict and infer such skills (Cook et al., 2011; Bley-Vroman & Chaurdon, 1994; Henning, 1983). The premise of the EI approach is that as sentences become more complex, the learner must make use of his or her interlanguage in order to accurately reconstruct what is heard. Therefore those who can accurately repeat longer sentences have access to a larger bank of linguistic knowledge and competence (metaphorically speaking) and are identified as more advanced speakers of the language (Ellis, 2006; Erlam, 2006). Vinther (2002) illustrated this process with a five-step model. The test taker first listens to the sequence of sounds that comprise the prompt. Next, the test taker decodes the sequence of sounds into chunks of meaningful linguistic units and stores the information in short-term memory. The test taker's familiarity with the linguistic system (grammar, vocabulary, context, etc.) dictates how much of the information in the prompt sentence can be contained in a single chunk. The test taker then interprets the prompt by syntactically and semantically processing the chunks from the decoding

process. The test-taker then recalls the information and produces the sentence, utilizing his or her linguistic system to reconstruct the prompt.

Some scholars have claimed that EI measures nothing more than the ability for rote repetition through the working memory (McDade, Simpson, & Lamb, 1982). Cowan (1996) explained that working memory is the portion of the memory that temporarily stores information only relevant to accomplishing a current task. There is no question that working memory plays an important role in EI tasks (Doughty & Long, 2003); however, there is still discussion in the literature about the degree of overlap between working memory and linguistic ability in EI. Erlam (2006) summarized the literature in this regard, providing three points of evidence that EI measures more than the ability to perform rote imitations. First, research has shown that working memory capacity is determined by the information in the learner's long-term memory (Baddeley, Gathercole, & Papagno, 1998). Next, Potter and Lombardi (1990) provided evidence that memory for the meaning of an utterance is retained longer than the memory for the form. Finally, Munnich, Flynn, and Martohardjono (1994) showed that sentences with incorrect grammar were corrected spontaneously during EI tasks, indicating that the learners weren't merely repeating what was heard based on working memory.

More recently, Okura and Lonsdale (2012) designed a study to measure participants' working memory abilities and their scores on an EI test in order to establish whether working memory ability had a significant impact on EI test performance. The participants were students at the English Language Center (ELC) at Brigham Young University (BYU). Each of the participants took a test designed to measure working memory and an English EI test used by the ELC. The correlation of the EI test performance and the working memory scores was insignificant (r=.249, p = .121). They reported, "the lack of significant correlations between

working memory and English EI scores…suggest that there is more to performance on EI tests than working memory capacity" (p. 2136). The literature gives ample evidence that EI measures implicit linguistic knowledge, and not just working memory ability.

Several studies have found success in comparing the results of EI tests to other language proficiency measures. Erlam (2009) conducted a study with 95 L2 learners of English from New Zealand and found a correlation of .87 between her EI instrument and the International English Language Testing System. Another study compared the use of a carefully constructed EI instrument with a more traditional speaking language achievement test (SLAT) and found a 0.74 correlation between the two tests (Graham et al., 2008.). Cook et al. (2011) compared the results of EI scores and OPI scores of 85 English as a Foreign Language learners in order to determine the predictive ability of the EI test. They used the EI scores to compute a predicted OPI score and found a 0.85 correlation between the predicted OPI scores and the actual scores. Along with these findings, many others have also reported significant positive correlations between EI performance and other measures of global language assessment (Call, 1985; Clay, 1971; Perkins, Brutten, & Angelis, 1986).

However, an important consideration in EI assessment is the development of the levels of the proficiency scale used in the creation of EI items. The proficiency scale used in this study was the ACTFL Proficiency Guidelines (1982). The ACTFL Proficiency Guidelines are based on a rating scale originally developed by the Foreign Service Institute of the U.S. Department of State in the 1950s. They were created to adapt this scale for use in schools and colleges. The guidelines for oral proficiency include 10 levels from novice low to superior. For a more complete description of the ten levels, refer to Breiner-Sanders, Lowe, Jr., Miles, & Swender (2000).

Although the ACTLF Proficiency Guidelines have been in use across the country for decades, the scale is not without its fair share of criticisms. In her survey of the literature, Liskin-Gasparro (2003) listed several of the criticisms that the ACTFL rating scale has received over the years. First, critics have pointed out that the guidelines for the scale were based more on intuitive judgments rather than actual data, especially with the listening and reading scales that have been accused of being modifications of the speaking guidelines. Next, the proficiency levels have been accused of being circulatory in that the definition of the level is the ability of the person who is able to perform at that level. But the definition was only defined because there are those that are able to perform at that level. Another criticism is that the validity of the rating scale is called into question because of its reliance on native speakers' abilities as a criterion against which the performance of non-native speakers is measured. Notwithstanding the criticisms of the scale, the scale is still widely used, and some believe in its popularity (Liskin-Gasparro, 2003; Norris & Pfeiffer, 2003). We have chosen to use this scale in spite of its flaws because of its pervasiveness and the lack of a widely recognized suitable alternative.

Creating a Valid EI Test

In order to operationalize the proficiency levels listed above, we first reviewed what factors contribute to item complexity for an EI test. EI items must be carefully selected in order for the item to effectively measure a test-taker's interlanguage system. The literature has shown that there are several ways that EI test developers have selected items to be used in an EI instrument. One way is for a researcher to carefully construct sentences of various levels by paying particular attention to the factors that make the sentence difficult. The researcher can consult with feature lists that assign grammatical features to proficiency levels, and build items with certain features to make them correspond to different proficiency levels. The same can be

done in terms of lexical complexity and sentence length—constructing sentences of various levels of difficulty by including various levels of lexical complexity and length. This approach requires a highly specialized linguist to be able to create such items. The recent research has shown that this approach can be used to result in high correlations to other language proficiency measures (Graham et al. 2008; Graham, McGhee, & Millard, 2010). Another recent study, however, has claimed that one possible danger of this approach is that test takers find many of the prompts created by researchers unnatural, unauthentic, and awkward (Christensen, Hendrickson, & Lonsdale, 2010).

Another way to select items for an EI instrument that has received particular attention in recent studies is selecting items from a corpus of naturally occurring language. This approach is much more systematic than the previously mentioned method of constructing each item. Because large language corpora exist for most of the world's major languages, this approach can be advantageous by placing the burden of item selection on these corpus tools rather than on an individual researcher. Items stemming from naturally occurring language have been shown to have high correlations with other proficiency measures, such as the OPI and Second Language Acquisition and Teaching certification. The study mentioned above claimed to achieve a 0.75 correlation, significantly better ($p < .05$) than the previous EI test's [speaking language achievement test] correlation ($r = 0.41$). Millard (2011) created an EI test for learners of French using the GigaWord corpus, a large corpus of naturally occurring written and spoken language. He administered his test to 94 participants and found a .92 correlation between his instrument and the OPI in terms of its ability to distinguish between levels of language proficiency.

Graham et al. (2008) found that the highest determiner of item difficulty was the item length in terms of syllables. The suitable length in syllables for items in an EI instrument

depends on the morphosyntactic features of the language. Miller (1956) has shown that the average individual is able to store about seven (plus or minus two) unrelated items at once in the working memory. Several more recent studies have suggested that four (plus or minus one) is a better representation of the working memory's capacity (Cowan, 2001). This research suggests that the length of items in an EI test should at least be greater than the working memory capacity limit in order to measure interlanguage ability. The range of sentence length for English language learners is between 6 syllables and 19 syllables (Graham et al., 2010; Vinther, 2002); however, the max number of syllables is higher in EI instruments that have been created for other languages. Millard (2011) found that the appropriate syllable range for learners of French was between 7 syllables and 25 syllables. Thompson (2013) found that the syllable range for learners of Spanish was between 7 and 34 syllables.

There are several reasons to suppose that the max length in terms of syllables for Russian will be longer than English. First, Russian is a highly inflected language, meaning that much of the grammar consists of adding affixes to the root of the word, which makes words several syllables longer. The assumption is that the affixes will be easier to chunk, meaning that a more proficient speaker of the language will be able to more easily chunk several syllables together because of the grammatical cohesion. Additionally, Russian has very few diphthongs. For example, the "-tion" morpheme in English has the equivalent of "ция" (tsee-ya) in Russian, which is two syllables in length. Many of the same words with the equivalent number morphemes have more syllables in Russian than in English. Again, the assumption is that chunking happens on a morphemic level rather than a syllable level, allowing native Russians to chunk morphemes of more syllables as easily as English speakers of less syllables (Bley-Vroman & Chaudron, 1994).

Research Questions

More research needs to be conducted to investigate the EI testing method's ability to infer oral language proficiency. Several of the studies above reported encouraging results regarding the usefulness of EI in L2 assessment, but most of them were studies involving English or Spanish as the L2. More research needs to be conducted with other language to investigate whether EI is a suitable approach for L2 acquisition in general. The research presented in this study answers these need by studying EI with a language (Russian) that has not been investigated in the literature and comparing the EI scores with Russian OPI scores. The research questions for this article are presented in the following bullet points.

- To what extent do the empirical Russian EI item difficulty levels align with their intended difficulty levels?
- To what extent does a criterion-referenced, proficiency-based EI test predict Russian language learners' OPI scores?

Methods

Research Context

The test created and evaluated in this study was given to students learning Russian in an intensive 9-week language learning program and students who have recently returned from an extensive experience abroad in a Russian-speaking country.

Test Design

In this section, we will explain the procedure used to create an EI instrument for learners of Russian. We will explain the procedure for extracting items, assigning items a difficulty score, and the initial process used to refine the item bank.

Item extraction. The items for this instrument came from two primary sources. The first came from the subcorpus of spoken Russian of the Russian National Corpus (http://ruscorpora.ru/en/search-spoken.html). The Russian National Corpus is a reference system based on an electronic collection of Russian texts. The subcorpus of Spoken Russian includes recordings of public and spontaneous Russian speech, including transcripts from Russian movies. This subcorpus is considered the best comprehensive source of naturally occurring Russian language. The corpus represents a well-balanced collection of speech that is situated in a large variety of contexts. The corpus includes nearly 150 million tokens taken from over 52,000 different sources.

The corpus was not available for download and was designed to only be searched and not browsed, so we developed a script to harvest the content. The script used the search parameters to look for all of the parts of speech. By searching for every part of speech, the results were able to access the entire corpus. Figure 1 shows a screenshot of the advanced search menu on the corpus. We simply checked every part of speech on the upper left box in order for the search to produce all of the content contained in the corpus. The script was written in Python. It systematically accessed each page of the results and scraped each of the results into a spreadsheet. The results were not already parsed into sentences. An additional script was required to detect a sentence ending punctuation mark. When such a mark was found, the script entered the following content as a new entry. After each of the sentences was entered as a separate entry, another script was written to detect the number of vowels in each sentence. The Russian language is such that the number of vowels in a word corresponds directly to the number of syllables. The results were presented in a spreadsheet where one column contained each of

the sentences scraped from the corpus and the other contained the number of syllables of each sentence.

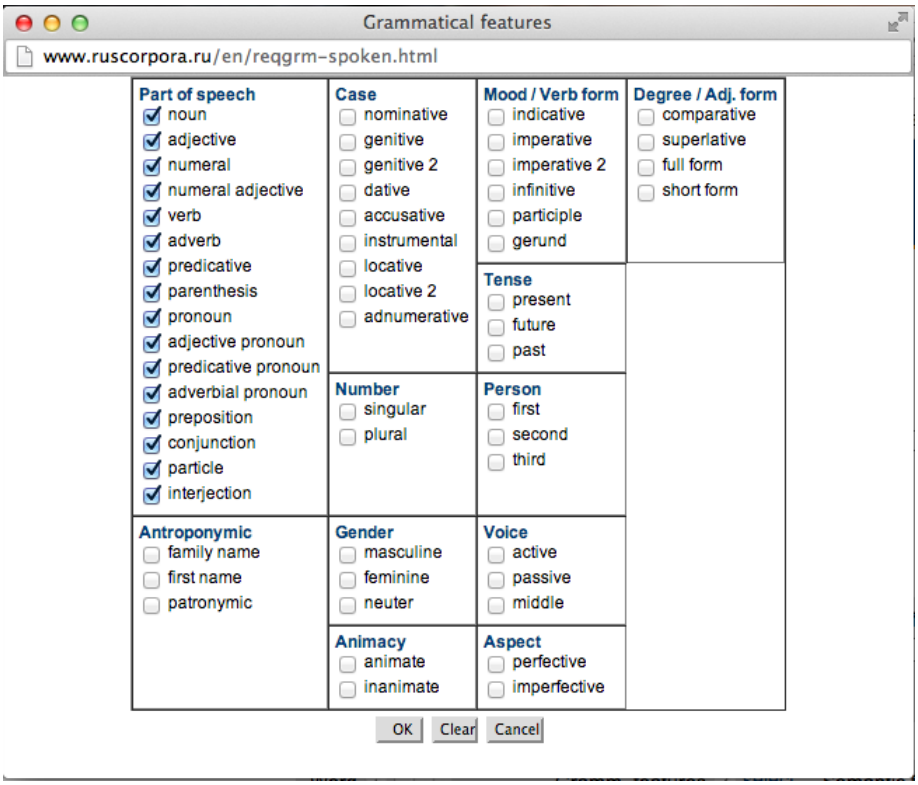

Figure 1. Screenshot of advanced search menu of the Russian National Corpus

The second bank of items was extracted from a social media website with personal stories and statements similar to the language that the learners would encounter in their experience abroad. The primary author copied the transcripts of 30 profiles into a document, which contained nearly 15,000 items. A similar script as was used to parse the data from the corpus was developed to parse the language in the document into individual sentences in a spreadsheet. Another script was created to count the number of syllables in each of the sentences, and the results were entered into a spreadsheet, where the first column contained the individual sentences

and the second column contained the number of syllables to the corresponding sentence. Upon completion of the extraction procedures, the banks from both sources were formatted exactly the same, and the processes described below were applied to each of the banks separately in a parallel manner.

Item complexity. After the item banks were created, the sentences were grouped in 3 levels according to levels 1-3 on the American Council on the Teaching of Foreign languages (ACTFL) scale (1 = intermediate, 2 = advanced, 3 = superior). We analyzed the sentences according to three factors: sentence length in terms of syllables, grammatical complexity, and lexical complexity. The purpose of determining the item complexity was an attempt at identifying the items that have the most discriminating power.

Sentence length. We conducted a pilot study to determine the max length in terms of syllables to be used in a Russian EI instrument. We created a bank of items that were similar in terms of grammatical and lexical difficulty. The items ranged from 26 syllables to 34 syllables in length. We then tested the items with 20 participants whose native language is Russian. We recorded a native Russian speaker reading each of the items, played the recording for the participant and asked the participants to repeat the item verbatim. 100% of the participants were able to accurately repeat the items that were 26 syllables in length. The average score for the items of 28 syllables was 93% with a standard deviation of 0.06. The average score dropped to 88% with a standard deviation of 0.11 for the items of 30 syllables. We found that native speakers of Russian struggled repeating back sentences that were longer than 30 syllables in length. This study suggests that a Russian language learner who is able to accurately repeat a sentence 30 syllables in length has reached a native-like performance for the instrument. We narrowed down the number of sentences in the item banks to those between 9-30 syllables in

length. We assigned all sentences from 9-15 syllables in length to the intermediate level, sentences from 16-22 syllables in length to the advanced level, and sentences 23-30 syllables in length to the superior level.

Grammatical complexity. In order to determine grammatical complexity, we used an indexed grammatical feature list created by OPI raters for Russian that outlines the grammar features that speakers of different proficiency levels have command over. We used this list to assign the grammar features a score corresponding to the level of difficulty from 1-3 (1 for intermediate, 2 for advanced and 3 for superior). See Appendix A for a list of Russian grammar features and their corresponding difficulty levels. These levels of difficulty correspond to the levels of difficulty used by the OPI and ACTFL raters: 0 = novice, 1 = intermediate, 2 = advanced, and 3 = superior. The definition of the novice level is the absence of language command indicated in the intermediate level, which is why we did not assign a score to items at a novice level. We then analyzed each sentence and marked the presence of each of the grammatical features by entering its score in separate columns in the spreadsheet. We computed the maximum score, which we used as the score to represent the level of difficult grammatical features in each sentence.

Lexical complexity. We used lemma frequency as the primary factor in determining lexical complexity. Lemma frequency is the cumulative frequency of all the word form frequencies of words within an inflectional paradigm. For example, although a verb may have several forms according to how it is conjugated, the lemma frequency couches each occurrence of the variation underneath the verb stem. This is important because we are not interested in the frequency of the variations of a word; rather we are interested in the frequency of the word and all of its forms. We used a lemmatizer tool developed by Serge Sharoff from the University of

Leeds to convert each of the word forms in the item banks to represent the lemma of the word (http://corpus.leeds.ac.uk/mocky/). Then we developed a script to search for the lemma word frequency of each of the words in the item banks, using a Russian lemma frequency list created by Serge Sharoff (http://www.artint.ru/projects/frqlist/frqlist-en.php). We assigned each item a lexical difficulty score, which equaled the score for the least frequent word in the item. According to the lexical difficulty score, we assigned the sentence a level from 1-3 on the ACTFL scale: items containing the most frequent 3,000 words were assigned level 1, items containing the words of frequency 3000-9000 were assigned level 2, and items containing words with frequencies above 9000 were assigned to level 3.

Item filtering. We extracted 20 items for each syllable length from 9-30 (440 items total: 220 from the Russian National Corpus and 220 from the social media website). The next step in determining which items to use in the EI instrument was to filter through the 440 items extracted and to identify the items with the most discriminating power based on the proficiency scale. This first step in filtering through these items was to simply throw out items that were assigned different levels according to the ACTFL scale for syllable length, grammar complexity, or lexical complexity. In other words, we only retained items that were on the same level in each of these three areas. The rationale for doing so is to increase control in the EI test. If an item is intermediate in terms of syllable length but superior in terms of grammatical complexity, then it becomes difficult to understand why the item did or did not perform well in the test. The purpose of this test is not to try and figure out which of these factors contributes to item difficulty and performance on the test. An additional filter we implemented was running the items by at least two native Russian speakers to have them eliminate items that are confusing and do not make sense taken out of context. We also removed items that contain large phrases that

will likely be chunked as an individual unit by most speakers of the language (for example, members of the LDS church will likely be able to chunk the phrase, "The Church of Jesus Christ of Latter-day Saints" as one unit).

The next process in filtering involved the expertise of two specialists trained in rating the OPI in Russian. These individuals went through the remaining items and confirmed the score assigned to them as mentioned above (1-3, intermediate to superior) according to the OPI rating standards. We discarded all of the items for which all of these three scores (grammar score mentioned above and the score of the two raters) did not agree. See Appendix B for the final list of items used in this EI test.

Test Administration Procedure

In this section, we will outline the procedure followed to test the items in the item bank to discover which items have the most discriminating power.

Participants. The participants for this study came from two groups. The first group of 52 were young men (28) and women (24) ages 18-26 learning Russian in an intensive program preparing them for experiences abroad in Russian-speaking countries. At the time of the study, these participants had been learning Russian for 4 to 8 weeks. The other 44 participants had recently returned from extensive experiences abroad from Russian-speaking countries. 11 were female and 33 were male ages 21-34. Three of the MTC employees were native Russian speakers.

EI test. The EI test consists of 72 items total split into 3 groups according to the ACTFL levels as discussed previously. The test has 24 items from each level chosen at random from the filtered item bank. The items were recorded by a male native speaker of Russian reciting each of the items at a normal speed with distinct, authentic, but not slurred or distorted pronunciation.

The browser-based administration program was able to pull from a database of pre-determined items to display to the test taker. The database included the audio track for each item, the item itself, and the item parsed into individual syllables. The display for the test taker was very simple, including a reference of how many items have been completed and how many remain (see Figure 2). The system randomly chose one of the 24 items per each level as a prompt for the test taker, and then the test taker repeated the prompt as accurately as possible. There was a delay of three seconds between each item, and then the system chose another item from the 24 items in that level and continued doing so until all 24 items from that level have been completed. The system then moved to the next level and repeated this process until all 72 items had been completed.

Administration. The EI test was administered in a computer lab with 12 computers. Before each session, the test was preloaded on each of the computers. The 54 students in the intensive program took the test in eight waves. The 44 students who had recently returned from being abroad took the EI test in 7 waves.

Scoring. This study will utilize the percentage scoring method to rate each of the items on the test. It is important to note that for this scoring method, individual syllables are the unit of measurement. If there is more than one mistake in a single syllable, the entire syllable will be counted as incorrect. If a syllable is missing, that entire syllable is incorrect. If a syllable or multiple syllables are correct but placed out of order, then only one of the syllables will be marked as incorrect.

The rater was presented with a clickable, parsed by syllable version of the item. The rater was also presented with a button that played the original audio recording and a button that played the recording from the test taker. The rater listened to the recordings, and clicked each syllable

of the item that was either pronounced incorrectly or was not pronounced at all. The rater ignored syllables that were repeated or inserted. Once satisfied that the item had been successfully rated, the rater moved on to the next item. This continued for each of the 72 items in the test. Two non-native Russian-speaking raters (who did not participate in the test) rated the EI tests. A third rater arbitrated any syllables that were not scored the same by the raters.

The percent score was then converted to a four-point rating scale where 0 indicates a score lower than a 10%, 1 indicates a score between 10%-50%, 2 indicates a score between 50%-90%, and 3 indicates the test-taker got higher than 90%. Within 2-3 days of taking the EI test, the test-takers took the computerized Oral Proficiency Interview (OPIc) in the same computer lab with the exception of 11 participants who had already taken the OPI within 3 months of taking the EI test. The OPIc is a test similar to the OPI that is designed to be administered online instead of in person. Instead of being interviewed by a live respondent, the test-taker is asked questions be a computer avatar and their responses are recorded and rated afterwards. Because of the inability for the computer to probe the test-taker to produce more advanced language, the OPIc is only able to assess language proficiency up to the advanced level.

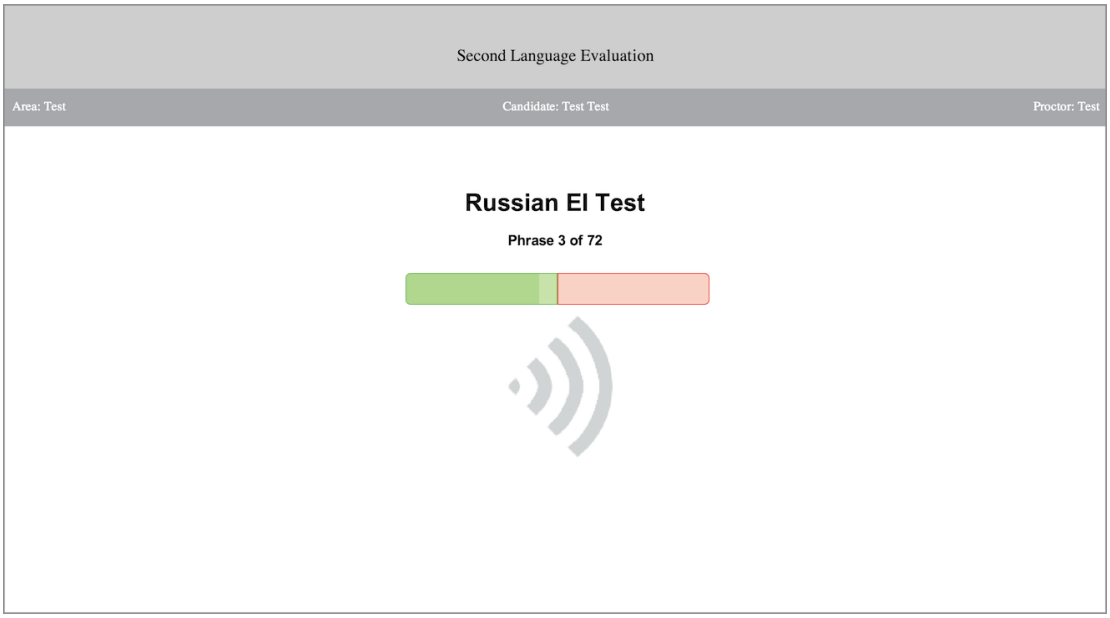

Figure 2. Test taker view of EI items

Results

In order to answer the questions in this study, we used the Rasch Item Response Theory (IRT) model to calculate the item difficulty statistics of the 72 items on the EI test. This analysis was accomplished through the Winsteps program (http://www.winsteps.com/index.htm). Because of its dichotomous and criterion-referenced nature, the Rasch measurement model is an appropriate form of analysis. For an in-depth history of the use of the Rasch model in language testing, see McNamara & Knoch (2012). Before reporting the findings for each of the research questions, we will present a diagnosis of the functionality of the rating scale followed by a reliability analysis of the test scores from the use of the scale.

Scale Diagnosis

The diagnosis indicates that the four-level scale mentioned above (0-3) functioned satisfactorily within the guidelines (Linacre, 2002). The average measures as well as the threshold estimates for each of the categories increased monotonically in each case. For each of the categories, the threshold estimates were between the recommended 1.4 to 5 logits between

each category, implying a distinction between each of the categories. Additionally, the spacing of the thresholds was regular, allowing the scale to be treated as interval data (see Figure 3). An examination of the category probability distributions showed that each category functioned well. The outfit statistics for the category ranged from 0.84 to 1.30, none of which were out of the acceptable range.

Reliability Analysis

The person ability estimates ranged from -5 to 9 on the scale with a mean of 0.18 (see Figure 4). Of the 96 people who took the exam, only two of the outfit mean squares exceeded 2.0, and the average for the set was 0.96. The internal separation reliability between the test takers was .99 with a separation strata index of 11.1. This value means we can be confident that the estimated person ability parameters indicate reliable differences between the performance on the EI test as defined by the four-point rating scale described above. The item ability estimates ranged from -7 to 6 on the scale with a mean of 0 (see Figure 5). The item separation reliability statistic was also .99, with a separate strata index of 9.92. The separate strata index for both person ability estimates and item ability estimates was higher than expected, and we verified the analysis to make sure this was not an error. We attribute the strength of the strata index to the wide range of proficiency levels of the learners and the three-level process we followed to determine item difficulty. Of the 72 items on the exam, only three of the outfit mean squares exceeded 2.0, and the average for the set was 1.05. These findings imply that the items were reliably distinct from each other and can easily represent at least 3 different difficulty levels that were intended.

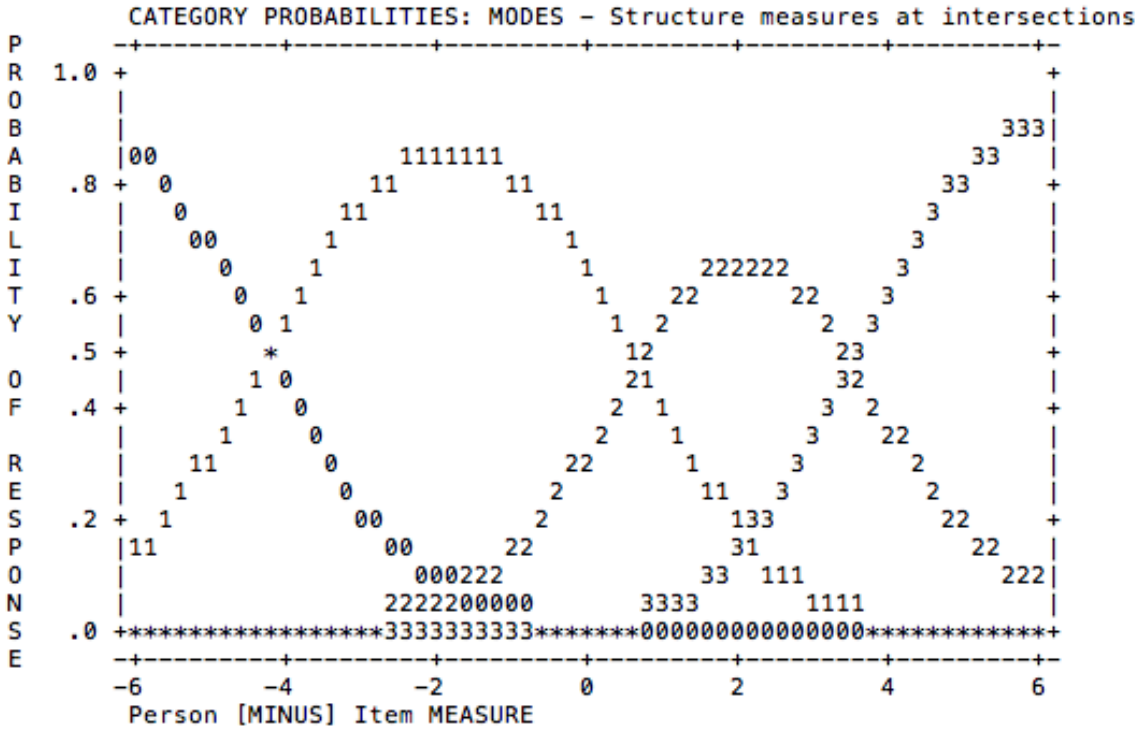

Figure 3. Russian EI rating category distribution

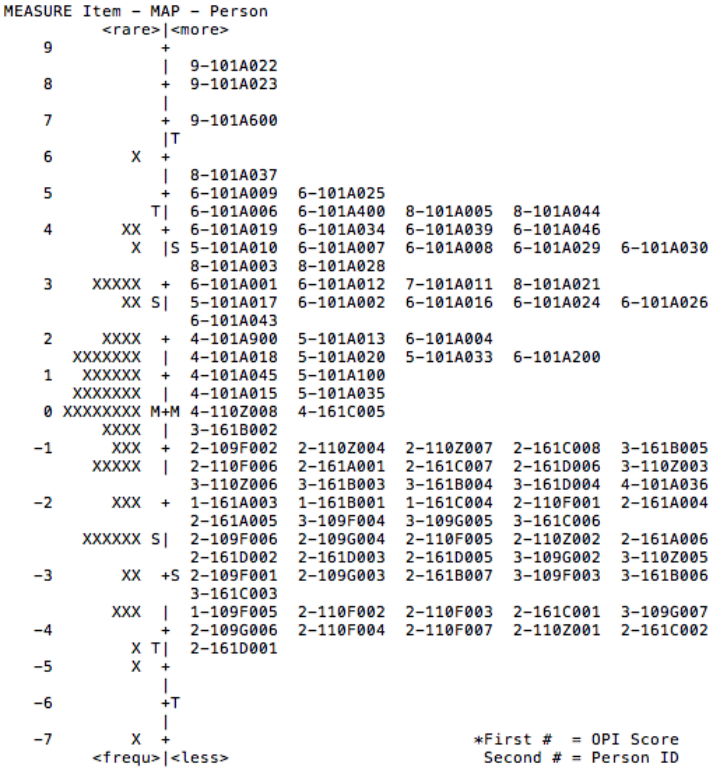

Figure 4. Russian EI person ability map

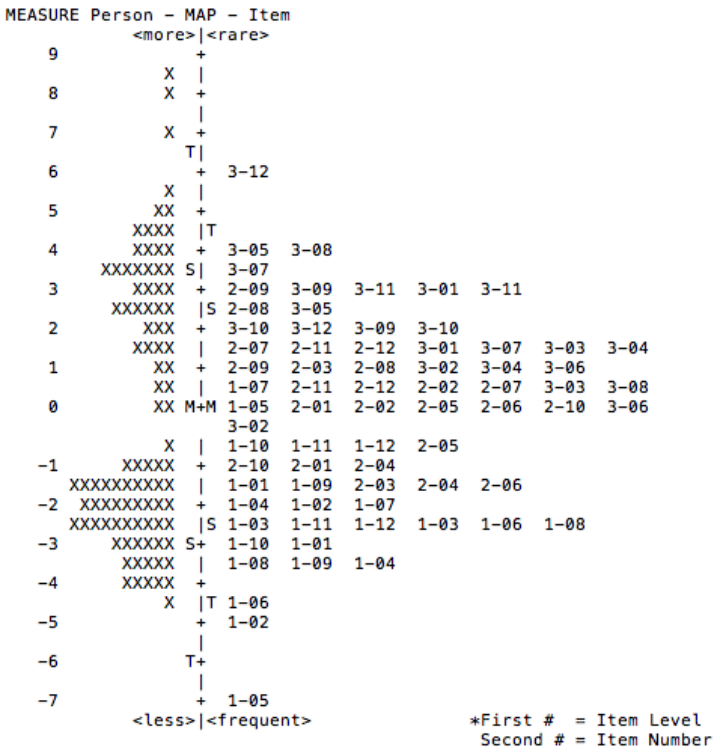

Figure 5. Russian EI item difficulty map

Question One: Alignment of Intended and Actual Item Difficulty Levels

A Pearson product-moment correlation coefficient was computed to assess the relationship between the item difficult logit measures and the intended ACTFL level for each item. The data passed the assumptions for using such a test in that the data are continuous and a scatterplot of the data affirm a linear relationship. There was a positive correlation between the two variables, $r = 0.773$, $n = 72$, $p < 0.001$. Increases in intended ACTFL level were correlated with increases in the item difficulty logit measure. Additionally, a one-way between subjects ANOVA was conducted to compare the effect of intended item ACTFL level (1-3) on the item's item difficulty logit measure. The data passed the assumptions for using an ANOVA test in that the logit measures were normally distributed with only a slight right skew with no extreme outliers. There was a significant effect of intended item ACTFL level on item difficulty logit measure at the $p < .05$ level for each of the three levels [F (2, 69) = 52.69, $p < 0.001$, $\eta^2 = .60$].

Post hoc comparisons using the Bonferroni test found statistical differences between intermediate (1) and advanced (2) items (mean difference = –2.38 logits, a 95% CI [–3.32, –1.41], and $p < 0.001$) and between advanced (2) and superior (3) items (mean difference = –1.60 logits, a 95% CI [–2.55 , –0.65], and $p < 0.001$). Taken together, these results suggest that the empirical difficulty levels as a whole align well with the intended item difficult ACTFL level. These data viewed in context of the first question of this study regarding the alignment of the empirical item difficulty measures with their intended difficulty levels indicate that the alignment is quite strong. However, a box plot of the data (See Figure 6) shows that for each level, there are some items that had item difficulty measures higher than the mean measure of the next intended ACTFL level.

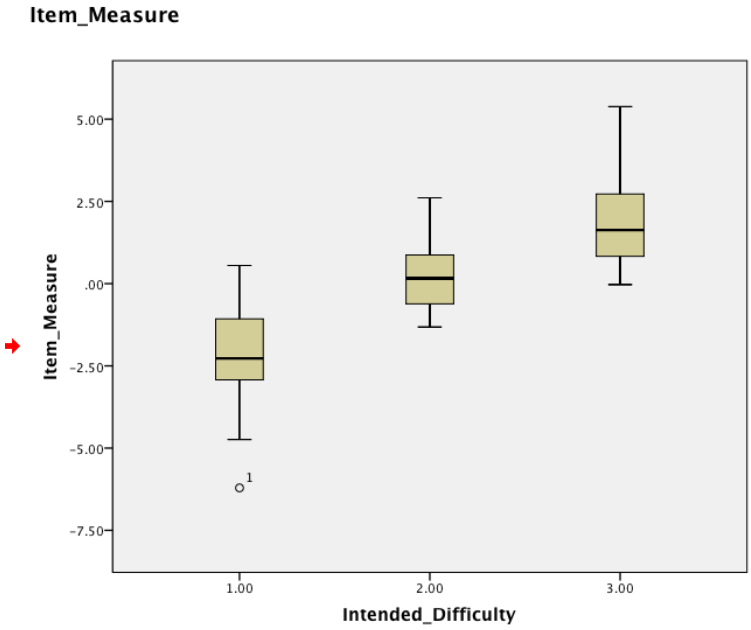

Figure 6. Boxplot of item difficulty statistics for intended difficulty

As seen above in the boxplot in Figure 7, there were several outliers in each group of intended difficulty. These items, their transliterations and translations are listed in Appendix C.

Two of the items were much easier than expected. Both of these items were intended to be level 3, but their item difficulty measures' placed them below the average of items in level 2. 1 of these items contained 23 syllables, and the other contained 24 syllables, both just above the cut point of 23 syllables to be in level 3. One of these items was incorrectly placed in terms of vocabulary difficulty. The item contained words within the 3,000 most common lemmas, which would make it level 1 according to lexical complexity. The other item contained only one word that placed it in level 3 according to lexical difficulty: "вдохновляющим" [vdakh-nav-lya-yu-shim] which means "inspiring." While this word may appropriately be infrequent in general speech, this word is much more frequent in the context for which the participants have learned Russian. Four of the items were much more difficult than intended. Two of the items were intended to be level one items but had item difficulty measures higher than the mean for the items in level two. These items both contained 12 syllables, approaching the limit to be considered level two items in terms of length. Both these sentences are highly marked. One contains nouns in three separate cases, while the other contains nouns in four cases. The other sentences in level one contain on average between one and two cases. This indicates that these sentences are grammatically dense. These factors may be the cause that they are more difficult than the others in the category. The other two items were intended to be level two, but had item difficulty measures higher that the mean for level three. These items both contained 22 syllables, just under the limit to be considered level three by length. Also, both of these items dealt with declining cardinal numbers, which is highly inflected and has many exceptions. Both of these factors may be a cause for their increased difficulty.

Question Two: Predictive Ability of EI Test for OPI scores

We used the Rasch IRT model to calculate the person ability estimates for the 96 participants in the study. The persona ability estimates were normally distributed and a scatter plot showed a strong linear relationship. Passing the assumptions, a simple linear regression analysis was conducted to find an equation to predict a subject's OPI score based on the person ability estimate of the criterion-references, proficiency-based EI test developed in this study. Subjects' OPI scores from the person ability estimate could be predicted by the following equation: y = .72x + 3.91, R^2 = 0.86, N = 96, r = 0.93. The scatterplot in Figure 7 below summarizes the results. These data viewed in context of the second question of this study indicate that the person ability measure is a strong predictor of a learners' oral proficiency as made evident by an OPI score. These data establish this EI test as a suitable testing instrument to indicate Russian oral language proficiency.

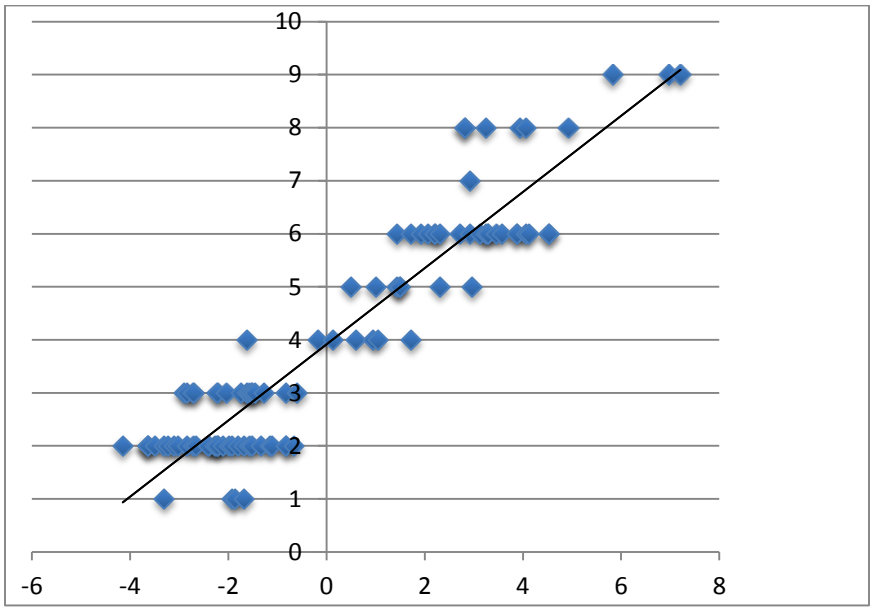

Figure 7. Scatterplot of person ability estimate and OPI score

Discussion

The relationship between the item difficulty measures and the intended difficulty level show a 77% correlation, and an ANOVA showed that the item difficulties of the items grouped by their intended difficulty levels were significantly different from each other with an effect size of 0.60, becoming more difficult as the intended level increased. This indicates that the items ascend hierarchically based on the ACTFL scale. In regards to question one of this study (How well do the intended item difficulty levels align with the actual levels?), these results provide good evidence that the item selection procedure proposed in the literature (Christensen et al. 2010; Millard & Lonsdale, 2011) and employed in this test was sufficient to produce an effective, predictive EI test, and the items performed as intended.

The regression analysis of the person ability estimates and the OPI scores ($R^2 = 0.86$, N = 96, $r = 0.93$) showed that the scores on the EI test strongly predicted the scores that the participants received on the OPI, providing important information for the second question of this research study (Can an EI test predict learners' OPI score?). While the EI test does not measure oral language proficiency, such a high correlation between the two tests suggests that one can with an acceptable degree of confidence infer oral language proficiency based on the scores of the EI test. As Erlam (2006) argued, there is strong evidence that EI measures an individual's interlanguage system and not just working memory ability. We suggest that EI is able to obtain such strong predictive power because EI is a measure of a learner's interlanguage system, and this system is at the root of oral language proficiency. These results are promising and support the results of other studies listed in the literature review of this article. Moulton (2012) conducted a similar study with ESL learners and found a strong correlation ($r = 0.83$) between her EI test and the Language Speaking Assessment (an assessment measuring oral proficiency

used at the Missionary Training Center in Provo, Utah). Millard and Lonsdale (2011) used a corpus as the source of his EI items, and in a similar comparison study found a strong correlation ($r = 0.92$) between the EI results and the OPI. The fact that there are several studies that have found such strong correlations, and few if any that have found contrary evidence adds to the validity of this field as a suitable option to indicate language proficiency.

Conclusion and Future Research

Although the results of this study are encouraging, there were several limiting factors that must be taken into account. While we have indicated the difference in levels for both the person ability estimates and the item difficulties, we have not shown that the person ability scores line up with the constructs of the item difficulties. For example, even though we have indicated which items are superior-level items and we have indicated which persons were superior-level persons, we have not provided evidence that these line up. Next, because of budget and scheduling constraints, the majority of participants in this study took the OPIc instead of the OPI. For the novice and intermediate levels, the OPIc is able to perform just as well as the OPI in differentiating between test-takers' ability (Kenyon & Malabonga, 2001). This is not the case for the advanced level. The upper level test-takers who took the OPIc and received an advanced score did not receive a delineation of low, mid, or high. On the 0-9 ACTFL scale from novice low to superior, those who received a score of advanced on the OPIc received a 6, which is the equivalent of advanced low. Although several of these test takers may have been able to receive a score of advanced mid, high, or even superior, the OPIc was not robust enough to differentiate at the higher levels. This lack of differentiating power hampered the ability of this study to differentiate among higher-level learners as well as it could differentiate among lower-level learners.

Additionally, we admit that the process of determining the complexity level of the items was somewhat arbitrary. More research needs to be done to determine how to more accurately determine the difficulty of each item. Further research is also needed to investigate whether the EI approach works for learners of Russian in a variety of contexts. Additionally, more research needs to be done to validate Millard and Lonsdale's (2011) success with using corpus tools as the source for effective EI items.

In spite of the limitations, this study still provides supporting evidence for the use of EI in language testing. The fact that the results in this study for a little-researched language (Russian) align with the results for studies of other prominent languages suggests that EI is not a language-specific phenomenon. More research is needed to investigate the utility of EI in more languages to confirm this. While this study does not attempt to identify which factors contribute to item complexity, controlling for sentence length, grammatical complexity, and lexical frequency was enough to produce strong results. Most importantly, this research suggests that EI can be used as a cheaper, faster alternative to the OPI and other expensive proficiency tests in order to surmise a learner's language proficiency.

References

American Council on the Teaching of Foreign Languages (1982). *ACTFL provisional proficiency guidelines.* Yonkers, NY: ACTFL

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105,* 158-173.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*, 355-371.

Bley-Vroman, R. & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E. E. Tarone, S. Gass & A. D. Cohen (Eds.), *Research methodology in second-language acquisition*. pp. 254-261. Hillsdale, NJ: Lawrence Erlbaum.

Breiner-Sanders, K., Lowe, Jr., P., Miles, J., & Swender, E. (2000). ACTFL proficiency guidelines—speaking. *Foreign Language Annals, 33* (1), 13-18.

Brown, J., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge University Press.

Call, M. (1985). Auditory short-term memory, listening comprehension, and the input hypothesis. *TESOL Quarterly*, *19*, 765-781.

Chaudron, C., Prior, M., & Kozok, U. (2005, July). Elicited imitation as an oral proficiency measure. Paper presented to the 14th World Congress of Applied Linguistics, Madison, WI, 2005.

Christensen, C., Hendrickson, R., & Lonsdale, D. (2010, May). Principled construction of elicited imitation tests. Paper presented to the Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC), 2010.

Clay, M. (1971). Sentence repetition: Elicited imitation of a controlled set of syntactic structures by four language groups. *Monographs of the Society for Research in Child Development*, *36* (3, Serial No. 143).

Cook, K., McGhee, J., & Lonsdale, D. (2011, June). Elicited imitation for prediction of OPI test scores. Paper presented to the Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, Portland, OR, 2011.

Cowan, N. (1996). Short-term memory, working memory, and their importance in language processing. *Topics in Language Disorders, 17,* 1-18.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87-185.

Doughty, C., & Long, M., (2003). *Handbook of second language acquisition.* Cambridge, England: Cambridge University Press.

Ellis, R. (2006). Modeling learning difficulty and second language proficiency: The differential contribution of implicit and explicit knowledge. *Applied Linguistics*, *27*, 431–463.

Erlam, R. (2006). *Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study.* Oxford: Oxford University Press.

Erlam, R. (2009). The elicited imitation test as a measure of implicit knowledge. In R. Ellis (Ed.), *Implicit and explicit knowledge in second language learning, testing and teaching*. Bristol, UK: Multilingual Matters.

Graham, R., Lonsdale, D., Kennington, C., Johnson, A., & McGhee, J. (2008). Elicited imitation as an oral proficiency measure with ASR scoring. In N. Calzolari (Conference Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, (Eds.), *Proceedings of the 6th International Language Resources and Evaluation Conference* (LREC'08), Marrakech, Morocco.

Graham, R., McGhee, J., & Millard, B. (2010). The role of lexical choice in elicited imitation item difficulty. In *Selected proceedings of the 2008 Second Language Research Forum*, (Ed.) Matthew T. Prior et al., 57-72. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2385.

Henning, G. (1983). Oral proficiency testing: Comparative validities of interviews, imitation, and completion methods. *Language Learning, 33* (3)*,* 315-332.

Linacre, J. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3* (1), 85-106.

Liskin-Gasparro, J. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals, 36* (4), 483-490.

McDade, H., Simpson, M., & Lamb, D. (1982). The use of elicited imitation as a measure of expressive grammar: a question of validity. *Journal of Speech and Hearing Disorders, 47* (1), 19-24.

McNamara, T., & Knoch, U., (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*. *29* (4), 553-574.

Millard, B. & Lonsdale, D. (2011, March). Developing French sentences for use in French oral proficiency testing. Paper presented at the Linguistic Symposium on Romance Linguistics, University of Ottawa, Canada.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63* (2), 81-97.

Moulton, S. (2012). *Elicited imitation as a measure of oral language proficiency at the Missionary Training Center*. (Unpublished master's thesis). Brigham Young University.

Munnich, E., Flynn, S., & Martohardjono, G. (1994). Elicited imitation and grammaticality judgment tasks: What they measure and how they relate to each other. In E. Tarone, S. Gass, and A. Cohen (Eds.) *Research Methodology in Second-language Acquisition*. pp. 227-245. NJ: Lawrence Erlbaum.

Norris, J., & Pfeiffer, P. (2003). Exploring the uses and usefulness of ACTFL oral proficiency ratings and standards in college foreign language departments. *Foreign Language Annals, 36* (4), 572-581.

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

Perkins, K., Brutten, S., & Angelis, P. (1986). Derivational complexity and item difficulty in a sentence repetition task. *Language Learning, 36,* 125-141.

Potter, M., & Lombardi, L. (1990). Regeneration in short-term recall of sentences, *Journal of Memory and Language, 29,* 633-654.

Radloff, C. (1991). *Sentence repetition testing for studies of community bilingualism*. Dallas: Summer Institute of Linguistics.

Thompson, C. (2013). *The development and validation of a Spanish elicited imitation test of oral language proficiency for the Missionary Training Center*. (Unpublished doctoral dissertation). Brigham Young University.

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

Appendix A: Russian Grammar Features for Proficiency Levels

Proficiency Level	Grammar Feature
Intermediate	Gender and number agreement in high-frequency words
Intermediate	Verb control high frequency verbs
Intermediate	Past, present, future conjugation in high frequency words
Intermediate	Adjectives and adverbs
Intermediate	Relative pronouns
Intermediate	Simple conjunctions
Intermediate	Adverbial time words (then, tomorrow, in the morning)
Intermediate	Ordinal numbers 1-100 in
Intermediate	Basic modal verbs
Intermediate	Impersonal constructions
Advanced	Passive voice
Advanced	Aspect
Advanced	Reflexive
Advanced	Prefixes of motion verbs
Advanced	Relative clauses
Advanced	Verb control
Advanced	Declensions of number in all cases
Advanced	Conditional
Advanced	Comparative adjectives
Advanced	Declension of proper nouns
Advanced	Definite pronouns
Advanced	Indirect speech
Superior	Participle constructions
Superior	Subordinate clauses of concession/compromise
Superior	Diminutive/affectionate nouns and adjectives

Appendix B: Items In the Russian Elicited Imitation Test

Item #	Syllables	Item
1	10	У меня пять прекрасных дочерей
2	10	Я знаю, что это Его Церковь
3	10	Она ощутила истинный мир
4	12	В своей жизни я стараюсь служить другим
5	12	Я очень люблю это Евангелие
6	11	Сейчас у меня есть сильная вера
7	12	Я встала с колен со слезами на глазах
8	11	Я никогда не была так счастлива
9	13	Бог любит меня и слышит мои молитвы
10	14	это Здорово помогать людям верить в Бога
11	14	У меня есть разные обязанности в церкви
12	14	Молитва укрепляет мою веру в Христа
13	9	она уже почти не болит
14	9	Что будем покупать на рынке
15	9	ты сегодня ездила к Насте
16	10	Не знаю как но я тебя видел
17	10	Я еще не совсем с ума сошла
18	11	Я просто спросил как у тебя дела
19	11	ты же сказала что тебя не будет
20	11	Я скажу тебе ответ на твой вопрос
21	12	У нас там на даче прекрасная осень
22	13	Да в советское время такого не было

23	14	Я очень рад что вы все сегодня сюда пришли
24	14	У нас был здесь один маленький случайный концерт
25	15	В любом спорте я всегда играл под этим номером
26	17	Я женат уже на протяжении 18 лет
27	17	Церковь помогла мне стать более хорошим человеком
28	15	Мне нравится работать с молодежью в нашей Церкви
29	17	Как и у любой другой семьи, у нас есть свои трудности
30	20	Я знаю, что мой Отец на Небесах призвал меня к этой работе
31	21	Я стараюсь подавать пример чистой жизни и высоких нравственных норм
32	22	Я остаюсь дома с моими четырьмя замечательными малышами!
33	23	Многие члены Церкви помогали нам самыми различными способами
34	22	я прочитала Книгу Мормона первый раз когда училась в восьмом классе
35	22	Я просыпаюсь каждый день с миром и надеждой благодаря моей вере
36	23	Отказываясь от комплимента, вы отказываетесь от Божьих подарков
37	15	Я ещё точно не знаю во сколько я поеду
38	15	чем ты планируешь заняться во время отпуска
39	16	Он ждал меня у гостиницы где я остановился
40	17	Здесь он чувствовал себя очень спокойно и уверенно
41	17	давайте всё-таки вернёмся к более радостным вещам
42	18	Какие у вас возражения против этого термина
43	18	Мне бы хотелось сразу сделать небольшое замечание
44	20	К сожалению сегодня более ста детей не попали в списки
45	22	И мы работали с пяти утра до двух часов ночи следующих суток
46	22	Родились люди которые не знали никакого другого языка

47	23	Если никто не возмущается это еще не значит что все всем довольны
48	22	в итоге мы должны прийти к некоторым выводам и рекомендациям
49	23	Я люблю встречать новых людей и укреплять уже существующую дружбу
50	24	Одна из величайших драгоценностей в моей жизни - это моя сестра - близнец
51	24	Я вижу много благословений благодаря тому, что я в Церкви с четырех лет
52	24	Каждое утро я молюсь, прося о терпении в преодолении трудностей
53	24	Я провела большую часть моей взрослой жизни, служа подросткам в нашем приходе
54	24	Я очень люблю следовать вдохновляющим примерам людей, которых встречаю
55	27	У нас двое замечательных детей, которые не дают нам особенно расслабляться
56	29	Моя семья – самая большая радость в моей жизни и действительно благословение с Небес
57	30	Мы были благословлены тремя очаровательными дочками, которых мы просто обожаем
58	30	Оглядываясь назад я понимаю, что люди вне церкви часто были лучше и мудрее меня
59	30	Фактически, это – одна из величайших радостей жизни – непрерывно учиться и развиваться
60	30	Это – простой принцип, но моление – это то, что в любое время под силу любому человеку
61	23	В любом сообществе людей существуют проблемы охраны правопорядка
62	23	я очень рада что наконец-таки закончилось это долбаное лето
63	23	есть очень много детей-инвалидов, нуждающихся в приемных родителях
64	24	Это стало для меня самым потрясающим и непростым занятием в жизни
65	27	Отмечу что за последние пять лет увеличилось число часто болеющих школьников
66	26	пожалуйста припомните на президентских выборах за кого вы отдали свой голос
67	27	Папа будучи рыбаком стал бригадиром когда образовался колхоз в тридцатом году
68	29	Мы рады приветствовать вас сегодня на нашем празднике посвященном дню посёлка Белогорка
69	29	Если сейчас у вас это мнение поменялось то за кого бы вы сейчас проголосовали

70	29	в это мгновение слышу какой-то вопль и только потом понимаю что это мой собственный крик
71	30	Есть ли среди вас смельчаки которые не побоятся совершить со мной в такое путешествие
72	30	Защита поддерживает заявленное ходатайство о допросе указанного свидетеля

Appendix C: Item Outliers

Item 04

- В своей жизни я стараюсь служить другим.
- V svoye Zhizni ya starayus' sluzhit' drugim
- In my life, I try to serve others.

Item 07

- Я встала с колен со слезами на глазах.
- Ya vstala skolyen sa slyezami na glazakh
- I stood from my knees with tears in my eyes.

Item 32

- Я остаюсь дома с моими четырьмя замечательными малышами.
- Ya ostayus' doma smoyimi chetir'mya samyechatyel'nimi malishami
- I stay home with my four wonderful boys.

Item 45

- Мы работали с пяти утра до двух часов ночи следующих суток.
- Mi rabotali spiti utrpa do dvukh chasov nochi slyeduyushikh sutok
- We worked from five in the morning until two in the morning the next day.

Item 54

- Я очень люблю следовать вдохновляющим примерам людей, которых встречаю.
- Ya ochen lyublyu slyedovat' vdokhnovlyayuwhim primeram lyudyei, kotorikh vstrechayu
- I really like to follow the inspiring example of the people that I meet.

Item 62

- Я очень рад что наконец-таки закончилось это ужасное лето.
- Ya ochen rad shto nakonyets-taki zakonchilos' eto uzhasnoye lyeto
- I am very glad that finally this terrible summer.

Chapter 3: Article Two

Exploring Domain-General and Domain-Specific items of an Elicited Imitation Test

Jacob Burdis

Troy Cox

Jennifer Bown

Brigham Young University

ABSTRACT

Elicited imitation is a language assessment method that requires test-takers to repeat sentences of increasing difficulty in the target language. The accuracy at which test-takers are able to repeat more difficult sentences indicates the test-takers' language proficiency. However, in EI, the factors that render an item more complex than another have not been definitively identified. This study investigates the effect of general domain vs. specific domain items on item complexity and overall test performance. The study depicted in this paper was conducted at an intensive 9-week language training center with 54 students preparing for extended experiences abroad and 44 students who had recently returned from extensive experiences abroad in Russian-speaking countries. The EI instrument used in this study contains items pulled from a general corpus and a corpus of content specific to the experience abroad in order to investigate whether item difficulty and test performance is different between the two item banks. We found that the mean score for the content specific test ($\bar{x} = .51$) was significantly higher than the mean score for the general test ($\bar{x} = .44$, $p < 0.001$). Additionally, the item difficulties for the specific items were significantly less that the item difficulties for the general items ($p < 0.05$), indicating that the context of the EI items played a significant role in test performance.

The Effect of Content Familiarity on Elicited Imitation

In the last decade, research regarding the validity and utility of Elicited Imitation (EI) testing instruments has become more prevalent in language assessment literature. EI is a unique approach to language assessment that claims to predict a test-taker's language proficiency through a series of item repetition tasks. Test-takers are confronted with items in the target language and are immediately expected to repeat back what they heard as accurately as possible. The assumption is that the better someone is able to repeat back more complex items, the more proficient they are in the language. EI is an attractive option for language assessment because it is relatively fast, economical, and effective when compared to traditional proficiency assessments. Because the test is concerned only with the accuracy of the repetition, the scoring procedures for EI tests are drastically simpler than other proficiency tests. One must simply determine if the measuring unit (typically a syllable) was said correctly or not. Because of this simplicity, many researchers have found success in employing automated speech recognition (ASR) technology in the scoring of EI tests (Cook, McGhee, & Lonsdale, 2011; Graham, Lonsdale, Kennington, Johnson, & McGhee, 2008)—rendering the test even more economical and attractive.

The success of an EI instrument depends heavily on crafting or choosing items that appropriately discriminate between the proficiency levels of the test takers. It is a rather burdensome and tedious procedure to ensure that the items employed in an EI test will effectively discriminate reliably between test takers of various proficiency levels. As will be discussed more fully in this article, much research has been dedicated to identifying what factors are important when creating effective and reliable EI items. Several factors have been considered when analyzing what precisely makes an item more difficult than another, including

item length in terms of syllables, grammatical complexity, and lexical complexity. The purpose of this article is to investigate whether using general domain or specific domain items affects item difficulty and person ability estimates. An EI test was created with half of the items general in nature and the other half discipline specific in a context familiar to the test-takers (religious context). The performance on these two groups of items are compared and analyzed in order to provide more information regarding what factors contribute to item complexity in an EI instrument. If this does influence performance on an EI test, this will give reason for EI test creators to be more cautious when selecting the content of the items so as to not favor one group of test-takers over another.

Literature Review

The basic model of an EI instrument consists of a test with several items in the target language that gradually increase in difficulty. Test-takers listen to a recording of a native speaker of the target language reading the item prompt, and then immediately repeat back what they heard and understood as accurately as they are able (Chaudron, Priori, & Kozok, 2005). While on the surface, this may seem like a memory exercise, many studies in the literature have provided convincing evidence that EI actually measures test-takers' interlanguage system (Bley-Vroman & Chaurdon, 1994; Erlam, 2006). For a more complete discussion regarding what EI actually measures, consult Burdis (2014).

Item Complexity

An influential factor on the performance of an EI test is the complexity of each item. Many studies investigating item complexity have concluded that item length in terms of syllables is the most influential contributor to item complexity. A famous study conducted by Miller (1956) investigated the storage capacity of working memory and found that the average

individual is able to store seven (plus or minus two) unrelated items at once. This study suggests that an average person unfamiliar with a language could theoretically repeat an item up to nine syllables in length. Perkins, Brutten, and Angelis (1986) studied adult ESL learners' performance and found that the lower threshold that began to discriminate for language ability and not working memory was items of seven to eight syllables. They also reported that syllable length was the most robust determiner of item difficulty in their EI test. A more recent study suggested that four (plus or minus one) is a better representation of the working memory's capacity (Cowan, 2001). Thus, it is important to construct EI items above that threshold so that working memory is not the sole ability being measured. A study conducted by Hendrickson, Aitken, McGhee & Johnson (2010) investigated which features of an item account for the item difficulty in an EI test. They constructed a test with 60 items ranging from 3 to 33 syllables in length to be administered to 376 learners of English at the English Language Center (ELC) in Provo, Utah. They used a 44-feature model to investigate which features contributed most to item difficulty. A step-wise regression reported that the model accounted for 67% of the variability in the difficulty measure of each item, and that sentence syllable count was the greatest contributor to model accountability ($R^2 = .65$). The next closest feature only accounted for 1% of the variability in the model. This suggests that the length of the item had far more impact on item difficulty than any of the other morphosyntactic or lexical features.

Graham, McGhee, and Millard (2010) investigated the role and influence of four factors on item difficulty on an EI test. They created an EI instrument with 60 items varying in length from 4 to 19 syllables. They split the items into 30 groups of two according to sentence length in terms of syllables and the lemmatized frequency range of the words in the items. They administered the test to 81 learners of English at the ELC in Provo, Utah. A regression analysis

showed that sentence length, lexical frequency and lexical density were significant predictors of average score, and morphological complexity was not. A step-wise regression analysis with the three significant predictors mentioned above as the independent variables showed that the three factors accounted for a little more than 83% of the total variance in item difficulty. A majority of the variance (73%) was attributed to sentence length, 8% to lexical frequency and 2% to lexical density. They also reported that the effect of lexical frequency was only constant for items of 15 syllables or less. This study provides additional evidence that sentence length is the most influential factor on EI item difficulty, but that lexical factors also had a significant effect. They reported, "In spite of the overwhelming effects of sentence length on item difficulty, this study has shown that lexical difficulty needs to be taken into account when creating sentences for EI instruments" (p. 69). This study illustrates the need to further study the role of lexical factors in the creation of EI items.

Language for Specific Purposes Testing

Although the subject of language specific testing (LSP) is thoroughly discussed in second language testing research, this subject has not been adequately explored in the literature of EI. Two recent studies have briefly addressed the content domain of EI items in EI testing. Both of these studies were with participants learning highly specialized language. The researchers made the assumption that the LSP testing model was appropriate for their EI tests, though they gave no justification for this assumption (Moulton, 2012; Thompson, 2013). Below we will explore several studies that have investigated this topic in language testing.

Douglas (2001) provided a clear description of the distinction between general purpose language testing and LSP testing. He described that in general purpose testing, the content is derived from a theory of general language ability or acquisition. General purpose tests typically

measure cognitive constructs such as communicative language ability. On the other hand, in LSP testing the content is taken from specific analysis of the target language use (TLU). LSP tests typically measure specific performance of language to be used in target situations. This distinction made by Douglas and further researched in the articles below show that in language testing (including EI) care should be made to choose and defend why one approach was chosen over the other.

In a study conducted with grade school learners of English, Romhild, Kenyon, and MacGregor (2011) investigated the effect of domain-general and domain-specific linguistic knowledge in the assessment of academic English language proficiency through the ACCESS for English language learners test battery. While this test is claimed to measure academic language proficiency, the researches found that much of the test consisted of language specific to the academic content domain. They found that domain-general and domain-specific linguistic knowledge played a significant role and accounted for variance in the performance of participants on the ACCESS test. They also found that the variance attributed to each factor differed depending on the proficiency level of the learners. This indicates that in such a content-specific test, variance attributed to domain-general and domain-specific can be blurred. This illustrates the complexity that the factor of domain-specific vs. domain-general content has on language assessment.

Douglas and Selinker (1992) devised a study to investigate whether a field-specific test would be more useful than a general-purpose test in predicting field-specific performance. They conducted a study with 31 Chinese chemistry graduate students in which they administered three tests: a field-specific English test, a general-purpose English test, and a chemistry performance test. They had the raters of the chemistry performance test assess whether the test-taker was

ready to successfully enter the chemistry field and correlated that answer to the test-takers' performance on the two tests. They found a significant correlation ($r = 0.50$, $p < 0.01$) between the raters' recommendations and the test-takers' score on the field-specific English test, while there was no significant correlation ($r = 0.34$) with the general-purposes test. This study provides evidence that when predicting field-specific performance is the goal, field-specific tests are more useful. This leads one to ask whether the opposite is true—when general proficiency is the goal, are general-purpose tests more appropriate?

Douglas (2000) later discussed at length that LSP tests are contrived language use events meant only to measure the test taker's language ability for a specific purpose and knowledge of the specialist field—not as a measure of global language proficiency. He reported that the best use scenario for LSP tests are to indicate target language use. Similar to the conclusions made above in the study of Chinese chemistry graduate students, he claimed that a well-developed LSP test is effective in measuring domain specific and not domain general performance. He argued that while LSP tasks often have a high degree of situational authenticity, they often lack an adequate degree of interactional authenticity. Thus the assumption that using domain-specific items in an EI language test will predict global language proficiency stands on shaky ground.

A more recent study investigated the English oral proficiency of air traffic controllers across work-related testing tasks and non-specific English tasks on aviation (Moder & Halleck, 2009). They found that the majority of the air traffic controllers (64%) received operational or above scores on tasks that were directly related to their everyday work routine, but a small minority (14%) received the same score for tasks that were deemed common-occurrence tasks for air traffic controllers. An even smaller minority (7%) received the same score for tasks that were less common for air traffic controllers. These results suggest that the scores produced from

limiting a proficiency test to a context that is very familiar to the test-taker ought not be used to assume general proficiency. In this extreme example, if domain-specific tasks were solely used to indicate global language proficiency, more than 50% of the participants would be incorrectly assumed to have operational language ability.

To this effect, several researchers have voiced their concerns about language for specific purposes (LSP) testing, stating that it has not been shown to be any more valid than a general proficiency test (Davies, 2001; Honderich, 1995). Also, Elder (2001) has brought up that a serious challenge with LSP testing is defining and identifying which testing tasks actually represent field-specific content, and which do not. She mentioned that the line is not so distinct as to be confident that an LSP test is equally familiar to all of the test-takers. Similar to the study done by Romhild et. al (2011), it is very difficult to control for the variation of domain-specific linguistic experience in the test takers. Because EI does not directly measure language proficiency but infers it, research is needed to explore how the issues of LSP testing affect this unique test design.

It is evident that more research is needed to identify the factors that contribute to the creation of EI items that effectively discriminate between language proficiency levels. Although item length has been shown in several studies to be the most influential factor in item complexity, Graham et al. (2010) have argued the need to further research the role of lexical factors on item complexity. Additionally, research has shown that in many facets of language learning—especially in language testing—LSP is an influential factor. We have not found research that has chosen to study the use of domain specific vs domain general items as a factor in item complexity for an EI test. This study will investigate this factor affects the difficulty of the items and the performance of the test-takers. Knowing whether the domain of the items

influences EI performance is crucial for the EI test creation process. If researchers only pay attention to sentence length, then there may be confounding factors that result in artificially higher or lower scores for a certain group of test-takers.

Research Questions

In this study, we sought to answer the following research questions.

1. To what extent do the scores of an EI test using general knowledge items differ from an EI test using content-specific items familiar to the test-taker? How is this difference affected by the language learners' proficiency level?
2. Do items from a specific content domain differ in difficulty from items from a general context? How is this difference affected by the items' intended proficiency level?

Methods

Research Context

In order to analyze the difference between general and specific content, it is necessary to find a large group of participants that share a similar lexicon and context for learning the target language. The instrument in this study was designed for students learning Russian in an intensive 9-week language-learning program, preparing for missionary service abroad in a Russian-speaking country and other students who had recently returned from missionary service experiences abroad in a Russian-speaking country. The shared learning context among this group is religious language. The language-learning method implemented by this program is heavily contextual and task-based in that students are expected to learn the language needed specifically to succeed abroad. Students are expected to leave the program with a command of the core language needed to function when immersed in the target language.

Test Design

The EI test created in this study had 72 items total and was administered to 96 students. Thirty-six items came from the subcorpus of spoken Russian of the Russian National Corpus (http://ruscorpora.ru/en/search-spoken.html) while the other 36 items were extracted from a religious social media website with personal stories and statements similar to the language that the learners would encounter in their experiences abroad. The items had been previously administered and validated and were found to be highly reliable (Burdis et al., 2014).

Because this study focused on analyzing the effect of specific domain vs. general domain items, we strived to control for other factors known to influence EI test performance. Each of the item banks (general & familiar) were grouped into 3 levels according to levels 1-3 on the American Council on the Teaching of Foreign Languages (ACTFL) scale (1 = intermediate, 2 = advanced, 3 = superior). There was no group for level 0 (beginning) because our definition of beginning speech on the ACTFL scale is the absence of command of any aspect of the language. For each of the levels, sentences were selected that contained features of that level according to sentence length, grammatical complexity, and lexical frequency. If any one of the categories for an item did not fit within the constraints for that level, it was omitted. This process was equivalent for both the general and familiar item banks. Table 1 illustrates the constraints of each category of complexity for the levels listed above.

Table 1

Constraints of Item Complexity for ACTFL Levels 1-3

	Linguistic Features			Content	
ACTFL Level	Number of Syllables	Grammatical Complexity	Lexical Frequency (Lemma)	Secular	Religious
Intermediate	9-15	Command of Level 1 features	0-3,000	12	12
Advanced	16-22	Command of Level 2 features	3,000-9,000	12	12
Superior	23-30	Command of Level 3 features	9,000+	12	12

The participants' language proficiency was also measured as part of this study. Eighty-five of the 96 subjects took a computerized Oral Proficiency Interview (OPIc) for Russian within a week of taking the EI instrument. The OPIc is a nationally recognized oral language proficiency test developed by the American Council on the Teaching of Foreign Languages (ACTFL). It is the computerized version of the Oral Proficiency Interview (OPI), which is a similar test administered by a live rater over the telephone. The OPIc is administered by a computer avatar. The subjects' responses are recorded and then rated by a certified rater. The OPIc is considerably less expensive than the OPI, and it is much easier to administer—since it is administered online, there is no need to make an appointment with a live rater. The limitation of the OPIc vs. the OPI is that it is only able to rate test-takers up to the advanced level. Eleven of the participants had taken an OPI test within three months of taking the EI test, so their OPI scores were used in place of the OPIc score.

Test Administration

The following section describes the administration of the EI test for this study. We will briefly discuss the participants of the study, the administration procedures of the EI test, and the scoring procedures of the EI test and the OPIc test.

Participants. The participants for this study came from two groups. Fifty-two of them were young men (28) and women (24) ages 18-26 learning Russian in an intensive program, preparing for missionary service abroad in a Russian-speaking country. At the time of the study, these participants had studied Russian for 4 to 8 weeks. The other 44 participants had recently returned from missionary service experiences abroad in Russian-speaking countries. Eleven were female and 33 were male ages 21-34. Three of the participants were native Russian speakers.

Administration. The EI test was administered in November 2013 in a computer lab with 12 computers. Before each session, the test was preloaded on each of the computers. The 54 students in the intensive program took the test in eight waves. Within 2-3 days of taking the EI test, they took the OPIc test in the same lab. The 44 students who had recently returned from Russian-speaking countries took the EI test in seven waves. As mentioned above, 11 of them had already taken the OPI within three months. The remaining students took the OPIc in the lab within a week of taking the EI test.

Scoring. The OPIc tests were professionally rated, and two non-native Russian-speaking raters (who did not participate in the test) rated the EI tests. A third rater arbitrated any syllables that were not scored the same by the raters. The percent score was then converted to a four-point rating scale that had been previously validated (Burdis et al, 2014) where 0 indicates a score lower than a 10%, 1 indicates a score between 10%-50%, 2 indicates a score between 50%-90%, and 3 indicates the test-taker got higher than 90%. Both the item and the person separation statistic was .99, indicating strong internal reliability between both the test-takers and the items on the test. Figure 8 shows the item difficulty map for the items in the study. The first number represents the items' intended difficulty level and the second number is the item number.

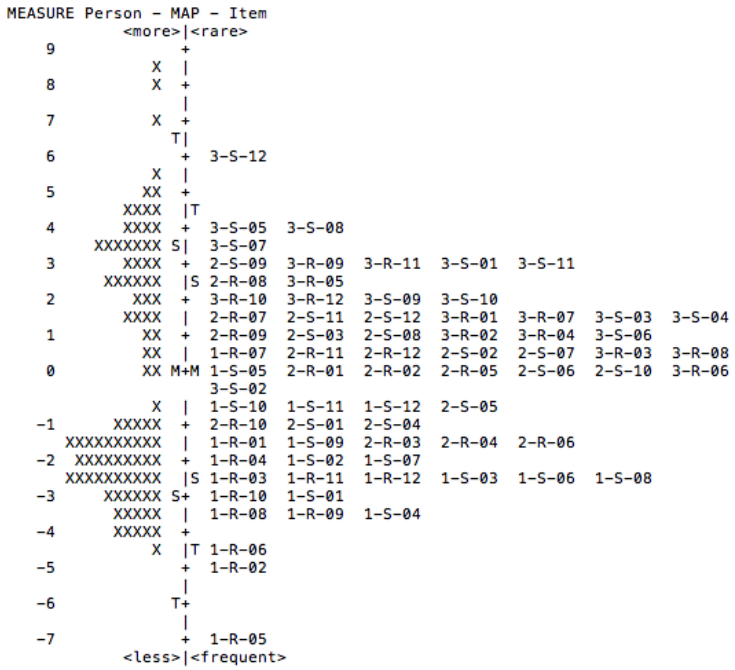

Figure 8. Russian EI item difficulty map

Results

Question One: Difference in Score for Specific vs. General EI Items.

Since a single test was administered to all participants, we used classical test theory to compare the effect of item type on the test score. The percent scores for the EI test were used to run a paired samples t-test to compare subjects' performance (percentage score) on familiar (religious) items vs. general items. There was a significant difference in the scores for religious ($\bar{x}$ = 0.51, SD = 0.25) versus general ($\bar{x}$ = 0.44, SD = 0.25) items; $t(95) = 21.08$, $p < 0.001$. See Figure 9 for a boxplot of the scores. These results indicated that the participants in this study as a whole performed better on the religious items vs. the general items. A one-way between subjects ANOVA was conducted to compare whether the subjects' difference in score for general and religious items varied depending on the subjects' ACTFL level as made evident by their OPI/OPIc score. There was not a significant variation of the subjects' difference in score

for general and religious items based on their ACTFL level at the $p < .05$ level for each of the three levels. These results suggest that although there is a difference as a whole between subjects' scores on the religious items vs. the general items, there is not enough evidence to support that this difference is any stronger or weaker for learners of a certain proficiency level.

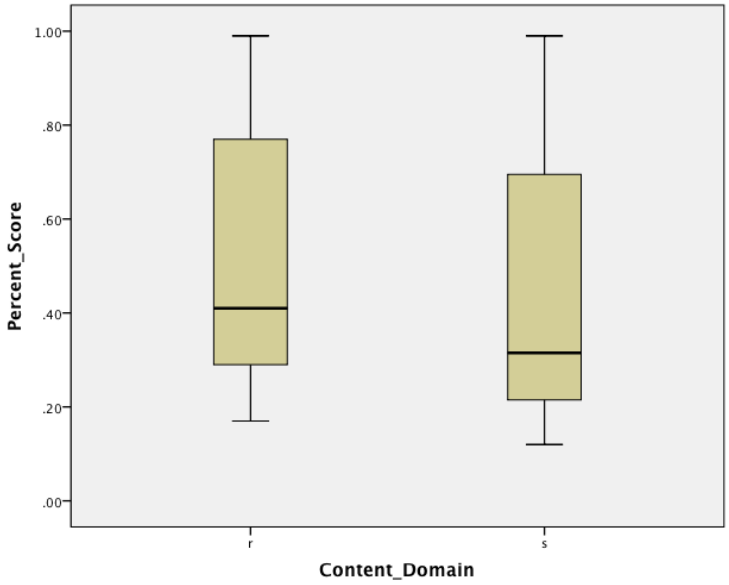

Figure 9. Boxplot of EI percent scores for religious and general items

Question Two: Difference in Item Difficulty for Specific vs. General EI Items

Since question two is centered on comparing the item difficulty measures of each item, the Rasch IRT model was used to conduct the analysis of item difficulty scores. A histogram of the data showed them to be normally distributed with a slight skew to the right and no extreme outliers, meetings the assumptions for using a one-way between subjects ANOVA. A one-way between subjects ANOVA was conducted to compare the effect of a generalized vs. specific content domain on the item difficulty measure for items on the EI test used in this study. Although the results are approaching significance, there was not a significant effect of content domain on item difficulty at the $p < .05$ level [F (1, 70) = 2.75, p = 0.113]. See Figure 10 for a

boxplot of the item difficulty measures. Although the difference in mean item difficulty score between the general ($\bar{x}$ = 0.40, *SD* = 2.03) and religious ($\bar{x}$ = –0.39, *SD* = 2.15) content groups was 0.79, these results indicate that this difference may be due to chance. However, according to the Rasch Model, a logit difference of 0.80 suggests that the students have a 30% probability of getting a general item correct and a 70% probability of getting a religious item correct.

We re-analyzed the religious items to identify any that were not overtly religious. We found 15 items (see Appendix A) that were not uniquely religious (e.g., "I have five wonderful daughters."). Since these items could be found in a general content domain, we hypothesized that they might be confounding the results and re-ran the analysis omitting these items. The items were evenly dispersed among the proficiency levels: five for the intermediate level, four for the advanced level, and six for the superior level. Because the items were evenly dispersed among the levels, removing the items should not affect the ability to compare with the general items. A one-way between subjects ANOVA was conducted to compare the effect of a generalized vs. the refined specific content domain on the item difficulty measure for items on the EI test used in this study. There was a significant effect of content domain on item difficulty at the $p < .05$ level [$F(1, 55) = 5.13$, $p = 0.028$]. See Figure 11 for a boxplot of the corrected item difficulty measures. These results suggest that the overtly religious items were easier that the general items. The mean logit difference between the religious ($\bar{x}$ = –0.93, *SD* = 2.31) and general ($\bar{x}$ = .40, SD = 2.03) content groups was 1.33, which suggests that students have between a 20-25% probability of getting a general item correct and a 75-80% probability of getting a religious item correct.

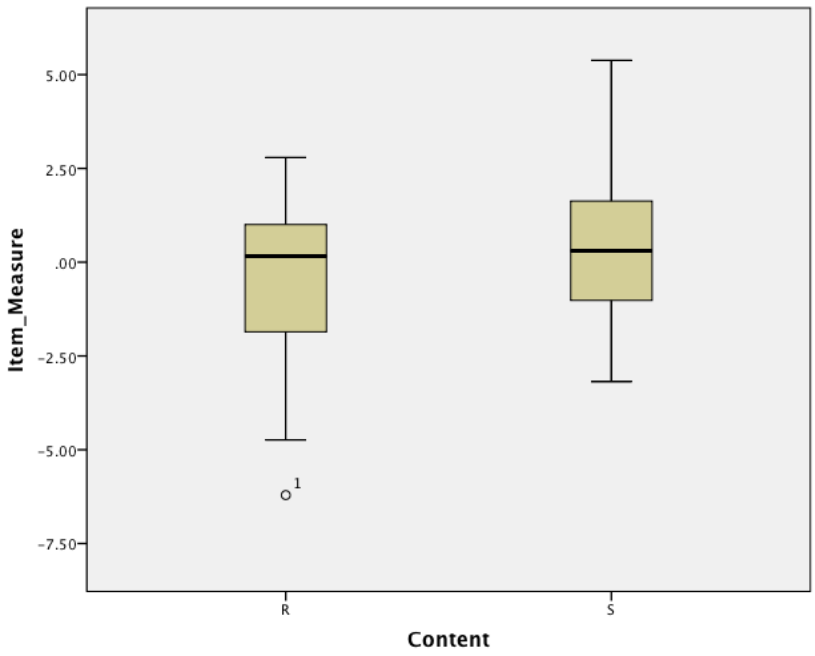

Figure 10. Boxplot of item difficulties for religious and general items

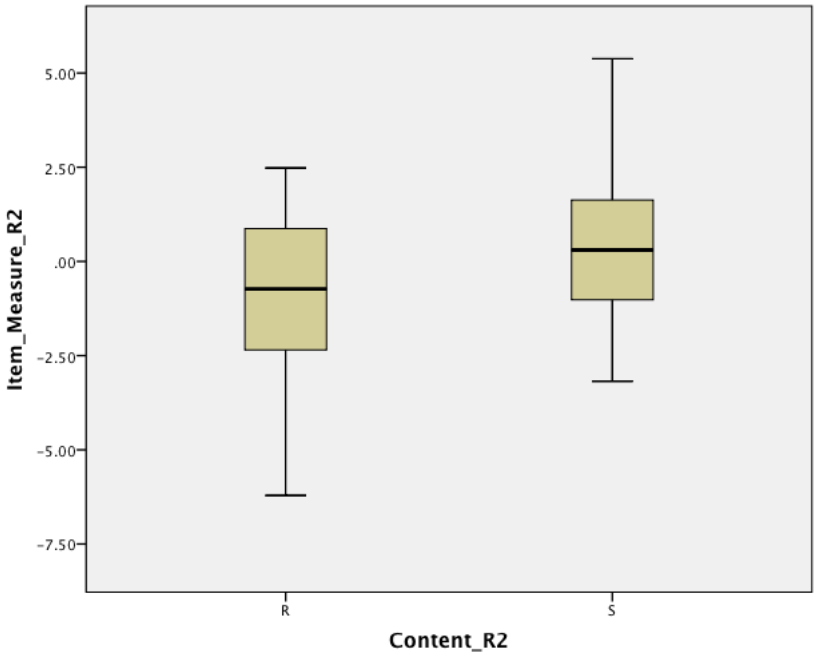

Figure 11. Boxplot of item difficulties for corrected religious and general items

A univariate ANOVA was conducted to compare whether the item difficulty measure for general vs. specific items varied based on the intended item ACTFL level of the items. As stated above, the item difficulty data showed a normal distribution slightly right skewed with no extreme outliers, meeting ANOVA assumptions. There was not a significant variation of the item difficulties for the two groups based on the intended ACTFL difficulty level of the items at the $p < .05$ level [F (2, 66) = 1.70, $p = 0.615$, $\eta^2 = 0.65$]. Although within each of the intended ACTFL groups the item difficulty for religious items was lower than for general items. These data suggest that there is no evidence that these differences are not due to chance. The analysis was repeated after omitting the 15 items that were determined to not be overtly religious. See Table 2 to compare the item difficulty scores. The univariate ANOVA found again that there still was not a significant variation at the $p < .05$ level [F(2, 51) = 1.44, $p = 0.224$, $\eta^2 = 0.73$]. But it is worth mentioning that removing the 15 items that were not overtly religious caused the p value to change by 0.391, indicating that removing these items moves the data closer to significance.

Table 2

Item Difficulty Scores Across Proficiency Levels

Intended Difficulty	ID Religious	ID Refined Religious	ID General
Intermediate	-2.60	-3.61	-1.63
Advanced	0.08	-0.16	0.44
Superior	1.34	1.17	2.37

Discussion

The data from this study indicate that the test-takers scored significantly higher on the religious items than the general items. As discussed, the participants in this study all learned Russian in a specific task-based learning context. The fact that the test-takers scored higher for the religious items vs. the general items indicates that EI testing follows general LSP testing trends. As discussed by Davies (2001) and Honderich (1995), LSP testing effectively predicts context-specific performance, but it has not been shown to predict general performance any better than general-domain language tests. As discussed above, Moder and Halleck (2009) showed that air traffic controllers had a high command of the language for tasks they engaged in every day, but a low command of more general English. They voiced their concern that LSP testing should not be used for establishing general proficiency for this reason. The findings of the current study suggest that the content domain of the items has a significant effect on item complexity of an EI instrument, adding to the argument that LSP testing for EI items may produce artificially high estimates of general language proficiency. It is interesting to note that significance was not found until problematic items were removed. These items were labeled religious items; however, there was little about them that distinguished them from the general items. Once they were removed from the comparison, significance was found. This gives further evidence that EI follows LSP testing in that an LSP EI test predicts specific performance better than general performance. It is interesting to note that removing the items that were not overtly religious lowered the *p* value by 0.391 and increased the effect size from $\eta^2 = 0.65$ to $\eta^2 = 0.73$. These data combined show that there is an effect of content domain of the items on person score; participants scored significantly higher on domain specific items than domain general items. If the purpose of EI testing is to infer general oral language proficiency, than any

assumption that an LSP EI test should be used for learners of specialized language is an argument that cannot be supported.

Conclusion and Future Study

These findings illustrate important implications for the creation of EI tests. While we do not know why the test-takers performed better on the religious items than the general items, we assume that at least one reason is because the test-takers were explicitly taught religious language content in the intensive language-training program. Even though it is doubtful that the learners had previously learned any of the exact sentences used in the religious EI test, we can assume that they were exposed to many of the words and phrases contained in them. This introduces a confounding factor that may indicate artificially increased test performance. Careful consideration must be made when analyzing the target audience of an EI test to ensure that the content of the test is not favoring one group, resulting in inflated test scores for that group because the content of the test catered more towards their context for language learning.

While these results are informative, a limiting factor of this study is that the number of items in each of the item banks was not sufficient. The analyses for both the person scores and the item difficulty showed that there was no significant difference in performance when taking into account the ACTFL level of the test-takers and the intended ACTFL levels of the items. There were only 12 items for each of the three ACTFL levels for the familiar and general item banks. To adequately see if content domain has more or less of an effect for different proficiency levels, a study will need to incorporate more items and more subjects, so that the individual groups are much higher.

Additionally, more should be done to control better for items containing words and contexts that are familiar to the test taker in order to truly study content domain for an EI

instrument. It is unclear whether the significant results came from specific training the learners received, or whether their familiarity with the topic in their native language made it easier for them to understand and produce the content in the target language. More research is needed to identify which of these factors contributing to content familiarity had a greater effect. Additionally, this study focused on religious vocabulary as the familiar content for test takers. This study should be replicated for other groups of language learners to try and obtain similar results.

The results of this study provide information that begins to show the importance of LSP testing for EI tests. We suggest that the same cautions and concerns that some researchers share for using LSP testing to predict general performance apply to using this model in EI testing. In order to use an EI test to predict general language proficiency, test-designers should be very cautious in choosing the contexts of the EI items. Limiting the contexts to a certain domain—especially if that domain is specific to the contexts in which the test-takers learned the language—may result in artificial results. The results of this study suggest that using such items is more appropriate for predicting context-specific language performance instead of general language proficiency. It is clear that although sentence length is a crucial determiner of item complexity, it is not the only one. Care must be made in regards to the content of the EI items when developing a valid EI test.

References

Bley-Vroman, R. & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E.E. Tarone, S. Gass & A.D. Cohen (Eds.), *Research methodology in second-language acquisition*. pp. 245-261. Hillsdale, NJ: Lawrence Erlbaum.

Brantmeier, C. (2005). Effects of reader's knowledge, text type, and test type on L1 and L2 reading comprehension. *Modern Language Journal*, 89 (1), 37-53.

Burdis, J. (2014). *Elicited imitation as a predictor of language proficiency for learners of Russian.* Unpublished manuscript.

Chaudron, C., Prior, M., & Kozok, U. (2005, July). Elicited imitation as an oral proficiency measure. Paper presented to the 14th World Congress of Applied Linguistics, Madison, WI, 2005.

Cook, K., McGhee, J., & Lonsdale, D. (2011, June). Elicited imitation for prediction of OPI test scores. Paper presented to the Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, Portland, OR, 2011.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing, 18* (2), 133-147.

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing, 18* (2), 171-185.

Douglas, D., & Selinker, L. (1992). Analyzing oral proficiency test performance in general and specific purpose contexts. *System, 20* (2), 317-328.

Elder, C. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing, 18* (2), 149-170.

Erlam, R. (2006). *Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study.* Oxford: Oxford University Press.

Graham, R., Lonsdale, D., Kennington, C., Johnson, A., & McGhee, J. (2008). Elicited imitation as an oral proficiency measure with ASR scoring. In N. Calzolari (Conference Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, (Eds.), *Proceedings of the 6th International Language Resources and Evaluation Conference* (LREC'08), Marrakech, Morocco.

Graham, R., McGhee, J., & Millard, B. (2010). The Role of Lexical Choice in Elicited Imitation Item Difficulty. In *Selected Proceedings of the 2008 Second Language Research Forum*, (Ed.) Matthew T. Prior et al., 57-72. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2385.

Hendrickson, R., Aitken, M., McGhee, J., & Johnson, A. (2010). What Makes an Item Difficult? A Syntactic, Lexical, and Morphological Study of Elicited Imitation Test Items. In *Selected Proceedings of the 2008 Second Language Research Forum*, (Ed.) Matthew T. Prior et al., 48-56. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2384.

Honderich, T. (1995). *The Oxford companion to philosophy.* Oxford: Oxford University Press.

Hudson, T. (1988). The effects of induced schemata on the "short-circuit" in L2 reading: Non-decoding factors in L2 reading performance. In P. L. Carrell, J. Decine, & D. E. Eskey (Eds.), *Interactive approaches to second language reading* (2nd ed., pp. 183-205). New York, NY: Cambridge University Press.

Johnson, P. (1982). Effects on reading comprehension of building background knowledge. *TESOL Quarterly*, 16 (4), 503-516.

Leeser, M. J. (2003). Second language comprehension and processing grammatical form: The effects of topic familiarity, mode, and pausing (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3* (1), 85-106.

Millard, B. & Lonsdale, D. (2011, March). Developing French sentences for use in French oral proficiency testing. Paper presented at the Linguistic Symposium on Romance Linguistics, University of Ottawa, Canada.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63 (2), 81-97.

Moder, C., & Halleck, G., (2009). Planes, politics, and oral proficiency: Testing international air traffic controllers. *Australian Review of Applied Linguistics, 32* (3), 1-16.

Moulton, S. (2012). *Elicited imitation as a measure of oral language proficiency at the Missionary Training Center*. (Unpublished master's thesis). Brigham Young University.

Perkins, K., Brutten, S. R., & Angelis, P. J. (1986). Derivational complexity and item difficulty in a sentence repetition task. *Language Learning, 36,* 125-141.

Recht, D., & Leslie, L. (1988). Effects of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology*, 80 (1), 16-20.

Romhild, A., Kenyon, D., & MacGregor, D. (2011). Exploring domain-general and domain-specific linguistic knowledge in the assessment of academic English language proficiency. *Language Assessment Quarterly, 8,* 213-228.

Thompson, C. (2013). *The development and validation of a Spanish elicited imitation test of oral language proficiency for the Missionary Training Center.* (Unpublished doctoral dissertation). Brigham Young University.

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

Appendix A: Items Not Overtly Religious

Item 01

- У меня пять прекрасных дочерей.
- U menya pyat’ prekrasnikh docherei
- I have five wonderful daughters.

Item 03

- Она ощутила истинный мир.
- Ona oshutila istini mir
- She felt true peace.

Item 04

- В своей жизни я стараюсь служить другим.
- V svoye Zhizni ya starayus’ sluzhit’ drugim
- In my life, I try to serve others.

Item 07

- Я встала с колен со слезами на глазах.
- Ya vstala skolyen sa slyezami na glazakh
- I stood from my knees with tears in my eyes.

Item 08

- Я никогда не был так счастлив.
- Ya nikogda nye buil tak shastliv
- I had never been so happy.

Item 24

- В любом спорте я всегда играл под этим номером.
- V lyubom sporte ya vsyegda igral pod etim nomerom
- In any sport I played under this number.

Item 25

- Я женат уже на протяжении 18 лет.
- Ya zhenat uzhe na protizheni vosyemnadsati lyet
- I’ve already been married for 18 years.

Item 29

- Как и у любой другой семьи, у нас есть свои трудности.
- Kak i u lyuboi drugoi sem'yi, u nas yest' svoi trudnosti
- Like any other family, we have our difficulties.

Item 32

- Я остаюсь дома с моими четырьмя замечательными малышами.
- Ya ostayus' doma smoimi chetir'mya samechatel'nimi malishami
- I stay home with my four amazing boys.

Item 49

- Я люблю встречать новых людей и укреплять уже существующую дружбу.
- Ya lyublyu vstrechat' novikh lyudyei I ukreplyat' uzhe sushestvuyushuyu druzhbu
- I love to meet new people and strengthen existing friendships.

Item 50

- Одна из величайших драгоценностей в моей жизни - это моя сестра – близнец.
- Odna iz vyelichaishikh dragotsyenostyei vmoyei zhizni – eto moya sestra – bliznets
- One of the greatest treasures in my life is my twin sister.

Item 55

- У нас двое замечательных детей, которые не дают нам особенно расслабляться.
- U nas dvoye zamyechatel'hikh dyetyei, kotoriye nye dayut nam osobyeno raslablyat'sya
- We have to amazing children that don't give us any time to relax.

Item 56

- Моя семья – самая большая радость в моей жизни и действительно благословение с Небес.
- Moya syemya – samaya bol'shaya radost' vmoyei zhizni I dyestvitel'no blagoclovyeniye snebes
- My family is the biggest joy in my life and is truly a blessing from heaven.

Item 57

- Мы были благословлены тремя очаровательными дочками, которых мы просто обожаем.
- Mui buili blagoslovlyeni tremya ocharovatyel'nimi dochkami, kotorikh mui prosto obozhayem

- We were blessed with three charming daughters whom we just adore.

Item 59

- Фактически, это – одна из величайших радостей жизни – непрерывно учиться и развиваться.
- Fakticheski, eto – odna iz byelichaishikh radostyei zhizni – nyeprerivno uchit’sya I rasvivat’sya
- In fact, continually learning and developing is one of the greatest joys of life.

Chapter 4: Conclusion

The results of the EI test created in this study exceeded expectations. The internal reliability reported for this instrument using the Rasch model was .99. This indicates that both the participants and the items in the test fit the model very well. We found that the EI test's ability to predict the test-takers' score on the OPI was very strong ($R^2 = .86$). These results have strong implications for the recognition of EI as a valid test to indicate language learners' oral proficiency. As EI gains more clout, we believe that there will be a shift in how large institutions, such as the MTC, approach proficiency testing. As EI instruments improve and become available in more languages, reliance on expensive and time consuming tests like the OPI will decrease.

We found that the test scores of content-familiar items ($\bar{x} = .51$) were significantly higher than the scores of general items ($\bar{x} = .44$, $p < 0.001$), and that that item difficulties for the content-familiar items were significantly less than the item difficulties for the general items ($p < .05$). This result is significant for the creators of EI tests. Much research has shown that sentence length is the primary factor that contributes to item difficulty. While this study does not refute that claim, it does add a caution that if an EI test creator only uses sentence length as the determiner of item complexity, confounding factors may produce inflated test scores. This study suggests that effort should be made to ensure that the content of the items on an EI test do not cater towards one group over another group.

As a whole, this study adds to the recent research validating EI as a valid, reliable language assessment option. EI has low face validity; many people find it hard to believe that a simple imitation can actually indicate oral language proficiency. More studies are needed to validate EI as a viable assessment instrument, showing that it correlates well with other standard

and accepted measures of language proficiency. Additionally, EI has only been studied for a handful of languages. The current study expands the research by investigating EI with Russian, which has received very little attention thus far in the literature. More research is needed to investigate EI with other languages. As EI is shown to work with more languages, the EI approach will become recognized as more of a general assessment approach instead of a language-specific assessment approach.

References—Articles One and Two

American Council on the Teaching of Foreign Languages (1982). *ACTFL provisional proficiency guidelines.* Yonkers, NY: ACTFL

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105,* 158-173.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*, 355-371.

Bley-Vroman, R. & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E. E. Tarone, S. Gass & A. D. Cohen (Eds.), *Research methodology in second-language acquisition*. pp. 254-261. Hillsdale, NJ: Lawrence Erlbaum.

Brantmeier, C. (2005). Effects of reader's knowledge, text type, and test type on L1 and L2 reading comprehension. *Modern Language Journal*, 89 (1), 37-53.

Breiner-Sanders, K., Lowe, Jr., P., Miles, J., & Swender, E. (2000). ACTFL proficiency guidelines—speaking. *Foreign Language Annals, 33, 1,* 13-18.

Brown, J., & Hudson, T. (2002). *Criterion-referenced language testing.* Caimbridge: Caimbridge University Press.

Burdis, J. (2014). *Elicited imitation as a predictor of language proficiency for learners of Russian.* Unpublished manuscript.

Call, M. (1985). Auditory short-term memory, listening comprehension, and the input hypothesis. *TESOL Quarterly*, *19*, 765-781.

Chaudron, C., Prior, M., & Kozok, U. (2005, July). Elicited imitation as an oral proficiency measure. Paper presented to the 14th World Congress of Applied Linguistics, Madison, WI, 2005.

Christensen, C., Hendrickson, R., & Lonsdale, D. (2010, May). Principled construction of elicited imitation tests. Paper presented to the Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC), 2010.

Clay, M. (1971). Sentence repetition: Elicited imitation of a controlled set of syntactic structures by four language groups. *Monographs of the Society for Research in Child Development*, *36* (3, Serial No. 143).

Cook, K., McGhee, J., & Lonsdale, D. (2011, June). Elicited imitation for prediction of OPI test scores. Paper presented to the Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications, Portland, OR, 2011.

Cowan, N. (1996). Short-term memory, working memory, and their importance in language processing. *Topics in Language Disorders, 17,* 1-18.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87-185.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing, 18* (2), 133-147.

Doughty, C., & Long, M., (2003). *Handbook of second language acquisition.* Cambridge, England: Cambridge University Press.

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing, 18* (2), 171-185.

Douglas, D., & Selinker, L. (1992). Analyzing oral proficiency test performance in general and specific purpose contexts. *System, 20* (2), 317-328.

Elder, C. (2001). Assessing the language proficiency of teachers: Are there any border controls? *Language Testing, 18* (2), 149-170.

Ellis, R. (2006). Modeling learning difficulty and second language proficiency: The differential contribution of implicit and explicit knowledge. *Applied Linguistics*, *27*, 431–463.

Erlam, R. (2006). *Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study.* Oxford: Oxford University Press.

Erlam, R. (2009). The elicited imitation test as a measure of implicit knowledge. In R. Ellis (Ed.), *Implicit and explicit knowledge in second language learning, testing and teaching*. Bristol, UK: Multilingual Matters.

Graham, R., Lonsdale, D., Kennington, C., Johnson, A., & McGhee, J. (2008). Elicited imitation as an oral proficiency measure with ASR scoring. In N. Calzolari (Conference Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, and D. Tapias, (Eds.), *Proceedings of the 6th International Language Resources and Evaluation Conference* (LREC'08), Marrakech, Morocco.

Graham, R., McGhee, J., & Millard, B. (2010). The role of lexical choice in elicited imitation item difficulty. In *Selected proceedings of the 2008 Second Language Research Forum*, (Ed.) Matthew T. Prior et al., 57-72. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2385.

Hendrickson, R., Aitken, M., McGhee, J., & Johnson, A. (2010). What Makes an Item Difficult? A Syntactic, Lexical, and Morphological Study of Elicited Imitation Test Items. In *Selected Proceedings of the 2008 Second Language Research Forum*, (Ed.) Matthew T. Prior et al., 48-56. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2384.

Henning, G. (1983). Oral proficiency testing: Comparative validities of interviews, imitation, and completion methods. *Language Learning, 33* (3)*,* 315-332.

Honderich, T. (1995). *The Oxford companion to philosophy*. Oxford: Oxford University Press.

Hudson, T. (1988). The effects of induced schemata on the "short-circuit" in L2 reading: Non-decoding factors in L2 reading performance. In P. L. Carrell, J. Decine, & D. E. Eskey (Eds.), Interactive approaches to second language reading (2nd ed., pp. 183-205). New York, NY: Cambridge University Press.

Johnson, P. (1982). Effects on reading comprehension of building background knowledge. *TESOL Quarterly*, 16 (4), 503-516.

Leeser, M. J. (2003). Second language comprehension and processing grammatical form: The effects of topic familiarity, mode, and pausing (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Linacre, J. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3* (1), 85-106.

Liskin-Gasparro, J. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals, 36* (4), 483-490.

McDade, H., Simpson, M., & Lamb, D. (1982). The use of elicited imitation as a measure of expressive grammar: a question of validity. *Journal of Speech and Hearing Disorders, 47* (1), 19-24.

McNamara, T., & Knoch, U., (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*. *29* (4), 553-574.

Millard, B. & Lonsdale, D. (2011, March). Developing French sentences for use in French oral proficiency testing. Paper presented at the Linguistic Symposium on Romance Linguistics, University of Ottawa, Canada.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63* (2), 81-97.

Moder, C., & Halleck, G., (2009). Planes, politics, and oral proficiency: Testing international air traffic controllers. *Australian Review of Applied Linguistics, 32* (3), 1-16.

Moulton, S. (2012). *Elicited imitation as a measure of oral language proficiency at the Missionary Training Center*. (Unpublished master's thesis). Brigham Young University.

Munnich, E., Flynn, S., & Martohardjono, G. (1994). Elicited imitation and grammaticality judgment tasks: What they measure and how they relate to each other. In E. Tarone, S. Gass, and A. Cohen (Eds.) *Research Methodology in Second-language Acquisition*. pp. 227-245. NJ: Lawrence Erlbaum.

Norris, J., & Pfeiffer, P. (2003). Exploring the uses and usefulness of ACTFL oral proficiency ratings and standards in college foreign language departments. *Foreign Language Annals, 36* (4), 572-581.

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

Perkins, K., Brutten, S., & Angelis, P. (1986). Derivational complexity and item difficulty in a sentence repetition task. *Language Learning, 36,* 125-141.

Potter, M., & Lombardi, L. (1990). Regeneration in short term recall of sentences, *Journal of Memory and Language, 29,* 633-654.

Radloff, C. (1991). *Sentence repetition testing for studies of community bilingualism*. Dallas: Summer Institute of Linguistics.

Recht, D., & Leslie, L. (1988). Effects of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology*, 80 (1), 16-20.

Romhild, A., Kenyon, D., & MacGregor, D. (2011). Exploring domain-general and domain-specific linguistic knowledge in the assessment of academic English language proficiency. *Language Assessment Quarterly, 8,* 213-228.

Thompson, C. (2013). *The development and validation of a Spanish elicited imitation test of oral language proficiency for the Missionary Training Center*. (Unpublished doctoral dissertation). Brigham Young University.

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