Exploring the Language of Assessment on Reading Proficiency Exams of Advanced Learners of Russian

Jeremy S. Evans
Brigham Young University - Provo

Follow this and additional works at: https://scholarsarchive.byu.edu/etd
Part of the Other Languages, Societies, and Cultures Commons

BYU ScholarsArchive Citation
Evans, Jeremy S., "Exploring the Language of Assessment on Reading Proficiency Exams of Advanced Learners of Russian" (2015). All Theses and Dissertations. 5651.
https://scholarsarchive.byu.edu/etd/5651

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

Exams of Advanced Learners of Russian

Jeremy S. Evans

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

Master of Arts

Troy L. Cox, Chair
Jennifer Bown
Teresa R. Bell

Center for Language Studies
Brigham Young University
December 2015

Copyright © 2015 Jeremy S. Evans
All Rights Reserved
ABSTRACT

Exploring the Language of Assessment on Reading Proficiency Exams of Advanced Learners of Russian

Jeremy S. Evans
Center for Language Studies, BYU
Master of Arts

Researchers have intermittently treated the topic of the language in which reading comprehension test questions should be presented in, or language of assessment (LoA). The overall consensus has been that questions in L1 lead to better scores and that questions in L1 should be used for reading comprehension particularly at the beginning levels. However, minimal research has been conducted at the advanced level, and no research has been found where proficiency items, empirically validated, were utilized in testing instruments. Furthermore, explanatory data from qualitative analysis has been sparse. The present research endeavored to satisfy these areas of needed research. It was found that a group of advanced learners of Russian performed better when MC questions were presented in English. Student attitudes, as revealed by survey items, depicted questions in L2 as more difficult. It was additionally found that vocabulary was one of the major factors in difficulty. Matters pertaining to validity and face validity surfaced and were designated, along with gathering other qualitative data, as the recommended direction of future research.

Keywords: reading comprehension, language of assessment, attitudes, language testing, advanced language learners, Russian
ACKNOWLEDGMENTS

I would like to give special thanks to Drs. Cox and Bown for being so patient and understanding throughout the whole adventure of writing this paper. They were also very willing to impart their expertise, counsel and kindness even when research was going slowly. An additional thank you to Dr. Bell who willingly joined my committee late in the process but nonetheless added to my success. To their praise, I felt like their colleague while at the same time recognizing my feeble grasp on topics of which they have such great mastery.

I would also like to express gratitude for BYU, including all of the faculty and staff, the facilities, the beautiful campus, etc. that played a valuable part in making my education and experience fulfilling and memorable.

A thank you is in order to my parents and family members for encouraging me to continue on and giving me moral and spiritual support on the hard days.

Finally, I would like to thank my Savior, Jesus the Christ who is lovingly involved in the details of my life, even the mundane and uneventful moments, many of which were spent working on research.
# TABLE OF CONTENTS

ABSTRACT ................................................................................................................................. ii

ACKNOWLEDGMENTS ............................................................................................................. iii

TABLE OF CONTENTS ............................................................................................................... iv

LIST OF TABLES ......................................................................................................................... vi

TABLE OF FIGURES .................................................................................................................. vii

Chapter 1: Introduction .............................................................................................................. 1

Chapter 2: Literature Review ..................................................................................................... 5

Reading Comprehension ........................................................................................................ 5

Reading proficiency .............................................................................................................. 6

Conceptualizing Previous LoA Research ............................................................................... 9

Level of Student ................................................................................................................ 10

Question Types ................................................................................................................ 10

The Effect of LoA on Scores ............................................................................................... 12

Attitudes and LoA ............................................................................................................ 19

Summary ............................................................................................................................. 21

Chapter 3: Methods ................................................................................................................... 22

Participants ........................................................................................................................ 22

The Instrument .................................................................................................................. 22

Survey instrument .......................................................................................................... 26

Elements of instrument not included in this paper ......................................................... 29

Procedure .......................................................................................................................... 29
LIST OF TABLES

Table 1 Previous LoA Research........................................................................................................... 14

Table 2 Numeric Means by Group ........................................................................................................ 31

Table 3 Attitudes Toward LoA ............................................................................................................. 35
| Figure 1. | Thematic Categories of Reading Proficiency | 8 |
| Figure 2. | Counterbalance Design of Reading Comprehension Exam | 24 |
| Figure 2. | Pyramid Diagram—Arrangement of Question Difficulty | 25 |
| Figure 2. | Screenshot—Passage and Questions | 26 |
| Figure 5. | Screenshots—Answer Change Surveys | 27 |
| Figure 6. | Screenshot—Confidence and Anxiety Measures | 28 |
| Figure 7. | Screenshot—Survey Items | 29 |
| Figure 8. | Mean Ability by Group and LoA | 32 |
| Figure 9. | Estimated Marginal Means | 33 |
Chapter 1: Introduction

When test developers are creating test and item specifications for foreign language (FL) reading comprehension assessments, a number of decisions must be made. For example, what item type(s) should be used? What types of reading passages should be used? How long should the passages be? How much time should examinees have to read the passage? What effect will the test have on the examinees taking it?

Among the many decisions to be made, the language in which second language (L2) test questions are posed, otherwise known as language of assessment (LoA), is still a developing topic in L2 reading assessment. In ESL settings, educators are prone to write exam questions in L2 because of the number of native languages represented among the students. The lack of feasibility in creating test questions in each student’s native language has deterred ESL researchers from investigating LoA issues in these settings. Foreign language and English as a Foreign Language (EFL) educators, on the other hand, have the perceived advantage, at least in this context, of a mostly homogenous group in terms of students’ L1. This offers flexibility in selecting the language in which the questions of a particular FL assessment should be presented. The opportunity to decide between L1 and L2 for test questions provides more latitude to educators, but it also generates additional questions. Out of these questions has sprung research pertaining to the effect of LoA, particularly as it relates to test scores (Brantmeier, 2006; Godev, Martinez-Gibson, & Toris, 2002; Gordon & Hanauer, 1995; Lee, 1986; Shohamy, 1984; Poh & Hock, 1979).

Shohamy’s 1984 study was one of the first to examine the effect of LoA on reading comprehension exams. Shohamy found that in addition to test method (i.e., multiple-choice, written recall, etc.), LoA likewise had a significant effect on students’ scores. In comparing
questions in L1 and L2 on open-ended and multiple-choice (MC) tasks, a hierarchy of difficulty was determined that would serve as a template for other LoA research. Open-ended questions in L2 were confirmed to be the most difficult to answer as reflected by lower scores, while multiple-choice questions in L1 were found to be the easiest. Thus, questions in L1 were easier between both question types. A hasty interpretation of these results may lead to the conclusion that questions in L1 should be used for assessing reading comprehension. Shohamy noted, however, that when “different scores are obtained on the same trait by using different testing methods [including LoA], the researcher can conclude that some of these scores are not a true assessment of the test taker’s knowledge of the trait” (p. 148). Therefore, difficulty alone is not the only factor involved in determining which language offers a “true assessment” of FL reading comprehension.

Since Shohamy’s study, LoA has received sparse treatment from researchers. While research has been conducted in limited areas (Brantmeier, 2006; Filipi, 2012; Godev, Martinez-Gibson, & Toris, 2002; Gordon & Hanauer, 1995; Lee, 1986; Nevo, 1989; Shohamy, 1984; Poh & Hock, 1979), many areas remain to be investigated including the effect of LoA on scores among advanced level language learners, affective elements, and questions that replicate previous research only with different languages as the L1 and L2. In order to truly understand the effect of LoA on test scores, supplemental measures in addition to statistical procedures have to be taken, such as assessing test takers’ attitudes. As it stands now, there are enough variables involved in L2 testing that statistically calculating whether questions in L1 or L2 lead to higher or lower test scores alone cannot give a full portrayal of the impact of LoA. There are several issues to consider when evaluating the existing research and drawing conclusions therefrom.

First, relatively little research has been conducted since the first LoA studies with the trend being to advocate using L1 for questions on L2 reading comprehension exams in foreign language
contexts. Second, replication of previous studies within different contexts is needed as well as to explore other relevant avenues related to LoA. In order to fully embrace the existing findings and understand more fully the issue of LoA more research will be needed to supplement and clarify previous research.

One of the important avenues that needs further investigation in this field of research is student proficiency level. To the present date, much of the L2 reading research has pertained to the beginning and intermediate levels of proficiency. After the handful of LoA studies among beginners, researchers and foreign language teachers seem to intuitively agree that the L1 should be used for testing reading comprehension at the beginning levels, believing that questions in the L2 would likely be beyond the students’ vocabulary and morphosyntactic knowledge. Only one known study has examined LoA among advanced learners at the collegiate level where it was found that when using written recall items, even certain advanced students scored lower in L2 (Brantmeier, 2006). The present study also endeavored to examine LoA among university students at the advanced level, only it utilized MC questions in order to better understand when and under what conditions LoA affects test scores. This information will also inform the discussion regarding the proficiency needed in order to be completely unaffected by LoA.

Second, there has been very little research into the attitudes of test takers in regards to LoA. A few known studies have assessed affective elements where LoA is one of the variables (Godev, Martinez-Gibson, & Toris, 2002; Filipi, 2012; Nevo, 1989). In her study, Filipi surveyed beginning and intermediate students of French and Japanese about their attitudes and preferences in regards to LoA on listening comprehension exams. Several aspects of listening comprehension exams distinguish them from reading comprehension exams such that they are different experiences for the test taker. On reading comprehension exams, for example, the test taker can revert back to the test repeatedly in order to understand and find answers, and
the pace of reading the text on a particular question is typically set by the test taker not the teacher. In contrast, learners taking L2 listening comprehension exams are restricted in efforts to revert back to and analyze the listening input by their own memory of what was heard, and they typically hear the input 2-3 times at a pace set by the teacher or the test recording. These and other differences suggest that listening as a skill is characteristically unique enough from reading that researchers would benefit from qualitative information regarding preference on the reading comprehension front where most empirical research lies.

**Purpose and Research Questions**

In view of the aforementioned opportunities for expanding the reach of previous research, the present study approached the issue of LoA in reading comprehension from the standpoint of reinforcing previous findings and investigating hypotheses about LoA at the advanced levels. In addition, the present paper endeavored to start a conversation about L2 readers’ attitudes toward LoA. The following are the research questions for the present study.

**Question 1:** What affect does language of assessment have on reading comprehension test scores among advanced learners of Russian?

**Question 2:** What are the attitudes of advanced learners of Russian toward language of assessment?
Chapter 2: Literature Review

In order to better understand the research questions, a brief background to reading comprehension is included. This background gives context to reading comprehension research, particularly testing in L2 settings. The previous research has explored the effect of LoA on scores of various question types and among differing populations, most of which are beginning and intermediate students. Correspondingly, advanced learners have received the least amount of attention. Regardless of the level of student, LoA studies that investigate affective variables such as attitudes have been few and thus additional attention to these variables would better inform the discussion of LoA.

Reading Comprehension

As a receptive skill, reading comprehension is an internal process dependent on many various sub-processes. For example, a series of studies highlights differences in internal processing when the purpose for reading is changed (Linderholm & van den Broek, 2002; Lorch, Lorch, & Klusewitz, 1993). When readers read with different goals in mind, such as for enjoyment, to learn, to evaluate, etc., they use different internal comprehension processes. Another example is that readers with little background knowledge use a different approach to reading a text than those who are already familiar with the topic of the text. Those with little background knowledge tend to exert more focus on details, instead of trying to gain more understanding of the context. On the other hand, readers with prior topic knowledge already have some context and therefore tend to be more selective with their attention. (Anderson, 1991; Brantmeier, 2005; Bugel, 1996; Shiotsu, 2007).

The numerous facets related to reading comprehension processing in L1 are complex, and reading in L2 adds an additional layer of complexity. For L2 researchers, the additional dimension and difficulty that are presented by L2 has led to many studies like those in L1 in
order to determine whether studies led to similar or different outcomes. Included among the topics from L1 research that are mirrored in L2 research settings are reading strategies, topic familiarity, comprehension exams, etc. In many cases, the research has revealed findings similar to findings in L1 studies.

Some of the most unique and difficult issues in reading comprehension that surface in L2 settings pertain to measuring it. Taking into consideration the fact that reading comprehension processes take place within the mind, educators are compelled to infer reading ability through external measures. In classroom settings this measurement takes the form of exams and assessments of varying kinds. The difficulty is that "by attempting to observe the reader's response we are bound in some way to affect that response" (Harrison & Dolan, 1979, p. 13). For this reason, the different types of examinations and question types have been the catalyst behind much debate and research, even without considering LoA. In L2 testing, however, LoA becomes another facet that can affect the test takers, their scores and attitudes.

**Reading proficiency.** In efforts to measure reading comprehension in a more uniform manner that offers educators a universal standard, proficiency guidelines of various kinds have been created. The most prevalent of the proficiency guidelines in academia today—ACTFL’s—stemmed from the ILR (Interagency Language Roundtable) proficiency rating system used by government agencies starting in the 1950s when government officials felt it exceedingly necessary to really know how proficient foreign language professionals in various agencies were. Today, ACTFL has become widely known among foreign language educators and researchers as a leader in providing and continuing to improve language standards and tests that can be incorporated in classroom testing, national testing, research, curriculum development, etc. The reach of the guidelines now encompasses the four major skills including reading. A general description of the reading guidelines from ACTFL (2012a) are as follows:
The ACTFL Proficiency Guidelines 2012 – Reading describe five major levels of proficiency: Distinguished, Superior, Advanced, Intermediate, and Novice. The description of each major level is representative of a specific range of abilities. Together these levels form a hierarchy in which each level subsumes all lower levels. The major levels Advanced, Intermediate, and Novice are divided into High, Mid, and Low sublevels. The subdivision of the Advanced level is new. This makes the Reading descriptions parallel to the other skill level descriptions. (p. 20)

Within these levels and sublevels, the goal is to find evidence of reading comprehension, or in other words, verify “the amount of information readers can retrieve from a text, and the inferences and connections that they can make within and across texts” (ACTFL, 2012a, p. 20). The manner in and extent to which L2 readers accomplish this goal is determinant upon the level. Beginners, for instance, are not expected to “retrieve” large amounts of information when their vocabulary is limited, or to make complicated inferences when their ability to express information in the L2 is still developing. Another developing aspect is the question of how beginners and their reading proficiency scores are effected when LoA becomes a factor. As mentioned above, beginners, arguably, should not be expected to take foreign language exams in the L2, especially if task information in L2 is beyond the capability of the student. However, as learners progress in L2 competencies including reading ability, they are presumably approaching a point where L2 ability equals or surpasses the level of the linguistic elements of questions in the L2. To complicate the issue, the reading tasks themselves are not static across levels. Correspondingly, the difficulty of the task changes and presumably so does the linguistic capability necessary to carry out the tasks in each level. A look at ACTFL’s reading proficiency familiarization manual gives representation of what and how readers must be able to perform across several categories of proficiency (Figure 1). As the function, content, context, text type,
vocabulary, grammar and culture facets change from intermediate to superior levels, so too do the tasks that assess proficiency at these levels. A simple example is the fact that intermediate readers are expected to read and understand high frequency vocabulary words, while advanced readers are expected to have a much broader vocabulary base that encompasses a wide range of topics. To be a superior level reader, the learner must have not only a breadth of vocabulary but a depth that includes specialized vocabulary. To some degree, then, this aspect of vocabulary knowledge ostensibly makes its way into the proficiency assessment, particularly when reading tasks are multiple-choice and much information in L2 must be read to answer questions. The other ACTFL categories of proficiency may or may not have as obvious of an impact as vocabulary on assessment but bringing attention to their possible role in assessments is valuable in the context of LoA.

![Figure 1. Thematic Categories of Reading Proficiency](image)

**ACTFL Proficiency Guidelines 2012 – Reading**

<table>
<thead>
<tr>
<th>Function</th>
<th>Content</th>
<th>Context</th>
<th>Text Type</th>
<th>Vocabulary</th>
<th>Grammar</th>
<th>Culture</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Superior</strong></td>
<td>Argumentation</td>
<td>Familiar and</td>
<td>Professional</td>
<td>Broad</td>
<td>Complex</td>
<td>Cultural references</td>
</tr>
<tr>
<td></td>
<td>Supported Opinion</td>
<td>unfamiliar abstract topics</td>
<td>Academic Literary</td>
<td>Precise</td>
<td>structures</td>
<td>Aesthetic properties</td>
</tr>
<tr>
<td></td>
<td>Hypothesis</td>
<td></td>
<td></td>
<td>Specialized</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Advanced</strong></td>
<td>Description</td>
<td>Concrete current</td>
<td>Public Education</td>
<td>Broad general</td>
<td>Sequencing</td>
<td>Most common cultural patterns</td>
</tr>
<tr>
<td></td>
<td>Narration</td>
<td>and general interest topics</td>
<td>Work News</td>
<td>vocabulary</td>
<td>Time frames</td>
<td>patterns</td>
</tr>
<tr>
<td></td>
<td>Exposition</td>
<td></td>
<td></td>
<td></td>
<td>Chronology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Explanation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Intermediate</strong></td>
<td>Convey basic information</td>
<td>Highly familiar</td>
<td>Highly familiar</td>
<td>Simple, predictable</td>
<td>Simple sentence</td>
<td>A few of the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>everyday content</td>
<td>everyday contexts</td>
<td>loosely-connected texts</td>
<td>patterns and strings</td>
<td>most common cultural patterns</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>of sentences</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 1. Thematic Categories of Reading Proficiency*

Note. Retrieved from ACTFL Reading Proficiency Familiarization Manual (ACTFL, 2012b)

**Criticism of ACTFL proficiency guidelines.** Most of the criticism of the ACTFL guidelines has pertained to speaking proficiency and measuring it via the Oral Proficiency
Interview (OPI). A major theme in feedback from L2 researchers related to the empirical validation of testing procedures such that one researcher critiqued that the foundational bases upon which most rating scales at the time were created were “intuitive judgments” (Fulcher, 1996, p. 208). Over the years, steps have been taken to improve and validate the speaking proficiency examinations and guidelines. Compared to speaking, the reading proficiency guidelines are much newer but have already benefited from empirical validation where aligned texts and reading tasks were used investigate the validity of the guidelines (Clifford & Cox, 2013). Clifford and Cox utilized data from three different studies and were subsequently able to affirm a significant difference between intermediate, advanced and superior levels.

Due to the fact that much previous research has taken place in various countries and at different ability and instructional levels over the last several decades, the research instruments have varied widely. Moreover, many studies were conducted before the surge of proficiency based guidelines, particularly for receptive skills such as reading. Much benefit could be gleaned from utilizing research instruments where the test items have been prepared and certified according to universal proficiency measures at a given level. Such questions, if validated for a given proficiency level, offer strength to the findings and confidence that items are relatively uniform in difficulty for that level. A universal set of guidelines can also assist test creators by offering a framework within which questions can be made appropriate to the level.

**Conceptualizing Previous LoA Research**

Language of assessment was first considered over 80 years ago by Stroebe (1930). Though it may now seem somewhat humorous, yet timelessly relevant to the modern language researcher, Stroebe began the conversation by stating that “comprehension tests with answers in English are a great convenience for the busy teacher” (p. 79). From then, it was not until approximately 50 years later that researchers first began a more intensive contemplation of LoA.
There have been two variables in the research design of prior studies of LoA’s effect on scores that have most largely dictated the direction of LoA research—level of student and question type.

**Level of Student.** Many have assumed that as students progress toward the advanced levels of proficiency, they will be less and less affected by LoA until there is no effect. This idea hinges in part on findings in previous research regarding the role of L1 in reading comprehension as students progress to advanced levels. For example, Upton (1997) and Upton and Thompson’s (2001) think-aloud protocols among L2 readers affirmed that L2 readers rely less on their L1 as they become more proficient in the target language (TL). Additionally, Kern (1994) found that L2 readers use less translation as a reading strategy at advanced levels of proficiency in comparison to the beginning stages. Discoveries from LoA studies have corroborated these findings. For example, Shohamy found that an advanced group of high school students was less affected by MC and open-ended questions in the L2 than the intermediate and beginning groups and that all groups performed better when questions were in L1. Shohamy and others have since concluded that lower test scores were likely due to issues related to vocabulary and wording of the question, as well as learner anxiety—a trait most prevalent among beginners.

**Question Types.** In LoA research three main types of questions used have been used—multiple-choice, open-ended and written recall. These questions types have emerged as a relevant variable in LoA research for several reasons. Foremost, the unique characteristics inherent in each question type carry certain ramifications depending on which language is chosen for LoA. Multiple-choice (MC) questions, for example, typically consist of a question stem followed by 3-5 answer options. When MC items are presented in L2 instead of L1, the test taker is required to sift through much more question-related information in the L2 in order to simply understand what is being asked and then to choose an answer. This has lead LoA researchers to
designate vocabulary and wording within the question as the major reason for lower scores when MC questions are in L2 (Alderson, 2000; Brantmeier, 2006; Shohamy, 1984).

By contrast, written recall items require very little of the test taker in terms of understanding the question. Typically, the written recall task is explained and instructions are given prior to the exam, therefore, no reliance on L1 or L2 is needed in order to understand the question. The students simply need to write down what they remember from the reading passage. A tremendous advantage of written recall items is not having question information that may cause misunderstanding. However, the requirement to write the answer presents a different set of challenges. Of greatest concern, weak writing ability could be mistaken for poor reading comprehension and vice versa. This dilemma is compounded when factoring LoA into writing tasks where L2 writing could be argued as the most difficult of the four skills in foreign language development. Open-ended tasks are typically somewhere between multiple-choice and written recall in regards to the demands they make on test takers.

**Problem solving vs. reading.** Another relevant aspect of question types that may have bearing on the effect of LoA relates to the types of strategies the test taker employs during examination. Rupp, Ferne, & Choi (2006) argue that MC items actuate a series of potential strategies that involve problem solving more than comprehending the text. In their study, Rupp et al. discovered that participants implement strategies that fall into two categories—unconditional and conditional. Unconditional strategies are those that test takers employ based on previous experience with MC items or for others reasons, and that are not related to characteristics of the passage or question. Conditional strategies are those that are selected based upon information gathered from the passage and question. Of the four unconditional strategies, two further subdivisions of participants were made: 1) those that read the passage or part of it first and then went back to the questions, and 2) those that read the questions first and then based on the
gathered information decided how to approach reading or scanning the passage. The four conditional strategies were also divided into two groups: 1) those that read or scanned all or part of the passage and then based on the information gathered decided on which strategy to employ next, and 2) those who read part or all of the questions and then based on that information decided which strategy to utilize.

Additional conclusions of made by Rupp et al. revealed several aspects of the MC test taking experience. First, when deciding which conditional strategies to use test takers based their decision on several perceptions about the passage and test item as it relates to difficulty. Factors such as length of text, and familiarity with vocabulary, the topic, and the text type were common pieces of information used in making estimates of difficulty. Second, test takers frequently resorted to underlining, highlighting and key word matching regardless of whether conditional or unconditional strategies were used. Third, the more difficult the text was perceived to be, the more test takers went back and forth between questions and texts engaging in a process of elimination until hopefully the correct response remained. The overarching criticism based on the findings and conclusions of Rupp et al. are that behavior exhibited by readers in non-testing contexts differs widely from that of MC testing contexts, and that MC items often evoke evidence of problem solving instead of the target task of reading.

The Effect of LoA on Scores

As mentioned above, other variables have proven important in determining the effect of LoA including the L1 and L2 used in testing. In addition, there are several different design-related variables incorporated by each LoA study that alter the way in which the results can be interpreted. These variables are instrument design, sample size, and statistical procedures used in computing results. Below is a review of LoA studies which includes details of each of the
pertinent variables in an attempt to give a clearer picture of the current state of LoA research (see Table 1). As such, the summaries of LoA studies are somewhat technical in presentation.

*Multiple-choice.* Prior to the Shohamy's study, and on a much smaller scale, Poh and Hock conducted a small-scale pilot study (1979). The 39 participants were explained as post fifth form learners of English in Malaysia whose native language was Bahasa Malaysia (fifth form is equivalent to senior year in high school). The effect of an immersion program was tested with a pretest / posttest study, and then the post test results were compared to the results of a delayed posttest that took place two days later. The results of the second part of the study offered insights into LoA by using questions in the L2 (English) for the posttest and questions in the L1 (Bahasa Malaysia) for the delayed posttest. The passage for the delayed posttest was identical to the passage used for the post test. LoA was the only change between the two exams. Poh and Hock found questions in L2 (English) significantly more difficult and thus lower scores than questions posed in L2 (English). They concluded that having questions in the L1 was indeed the best scenario, particularly for students “where English is a weak foreign language.”
Table 1

*Previous LoA Research*

<table>
<thead>
<tr>
<th>Researcher/Study Year</th>
<th>L1</th>
<th>L2</th>
<th>Level of Student</th>
<th>Participants</th>
<th>Question Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poh and Hock 1979</td>
<td>Malay</td>
<td>English</td>
<td>Post High School</td>
<td>Students described as post fifth form learners of English (n=39) (fifth form is equivalent to senior year in high school). These students were entering university.</td>
<td>MC</td>
</tr>
<tr>
<td>Shohamy 1984</td>
<td>Hebrew</td>
<td>English</td>
<td>Low, Int and High (High School)</td>
<td>Students in 12th grade in a high school setting in Israel (n=2000), and they were divided into proficiency groups of low, intermediate and high.</td>
<td>MC and Open ended</td>
</tr>
<tr>
<td>Lee 1986</td>
<td>English</td>
<td>Spanish</td>
<td>Beg and Int (University)</td>
<td>Students enrolled in four different semester level (n=320; 80 participants per level) . Spanish classes at Michigan State and University of Michigan for 1st and 2nd years.</td>
<td>Written Recall</td>
</tr>
<tr>
<td>Nevo 1989</td>
<td>Hebrew</td>
<td>French</td>
<td>Intermediate (High School)</td>
<td>Students in tenth grade in High School in Israel (n=42).</td>
<td>MC</td>
</tr>
<tr>
<td>Gordon 1995</td>
<td>Hebrew</td>
<td>English</td>
<td>10th Graders (High School)</td>
<td>Students in 10th grade (n=28) studying English in Israel.</td>
<td>Think-aloud protocols (MC and open ended)</td>
</tr>
<tr>
<td>Godev 2002</td>
<td>English</td>
<td>Spanish</td>
<td>Intermediate (University)</td>
<td>Students in third-semester (intermediate) of Spanish at a university (n=28).</td>
<td>Open ended- with different language for stem and answer</td>
</tr>
<tr>
<td>Brantmeier 2006</td>
<td>English</td>
<td>Spanish</td>
<td>Advanced (University)</td>
<td>Students enrolled in an advanced-level Spanish grammar and composition course at a private university in the Midwest (n=106).</td>
<td>Written Recall</td>
</tr>
</tbody>
</table>
While not a reading comprehension study, Filipi’s (2012) listening comprehension research utilized Item Response Theory (IRT) computation to capture the relationship between LoA and test scores from MC questions. It was found that, on the whole, LoA did not have any impact on the outcomes. There were a few questions, however, that asked for simple, explicitly stated information that turned out to be easier for examinees. This fact, stated Filipi, could be due to greater ease of copying and pasting answers when the questions require simple, explicitly stated information. The author concluded that “the language in which an item is written may be one factor in determining item difficulty, but that it may work with other factors in item design to make an item easier or more difficult” (pp. 525-526).

Multiple-choice vs open-ended. Shohamy’s initial study (1984) garnered recognition for several reasons. First, Shohamy incorporated a massive sample size of approximately 2000 participants—a rarity that serves as a great strength to the findings. Second, her study spanned across three different levels of proficiency: low, intermediate and high thus allowing for a more complete view of LoA and its effect on scores. Although, it must be mentioned that these levels are divisions within a high school setting and not necessarily divisions according to global proficiency scales. Third, the study compared two of the most favored question types for measuring reading comprehension—multiple-choice and open-ended. The findings revealed that regardless of the question type, the participants scored lower when the questions were in the L2, which in this experiment was English (L1 = Hebrew). Of the four variations of question type, open-ended in the L1, open-ended in the L2, MC in the L1 and MC in the L2, open-ended in the L2 proved to be the most difficult for the participants. The easiest question type, or the question type for which the highest scores were received was MC in the L1. When the results were subdivided by proficiency level, it was found that the scores of high-proficiency students on L2
questions were not significantly different from those in L1. In other words, LoA had no significant effect.

Gordon and Hanauer (1995) conducted study similar to that of Shohamy’s among 10th grade EFL students whose native language was also Hebrew, but he additionally investigated some qualitative elements that served to clarify the mental processes that occur during the different assessment tasks. Using the same question types of open-ended and MC in the L1 and L2 and the same questions for each participant with different variations of questions type and LoA, it was similarly found that the lowest scores were from the open-ended questions in the L2. On the other end of the continuum of student performance, the highest scores similarly came from MC questions in the L1. However, a think-aloud protocol was included that contributed to understanding the construction and development of the test takers' mental model under the conditions of the various assessment tasks. One major conclusion that emerged from the think-aloud tasks was that MC can offer the test taker more information possibly influence the ever evolving development of the mental model. Open-ended questions, per contra, provide less information to the test taker. In either scenario, the assessment task may be a source of information to the test taker. Gordon and Hanauer also noted that assessment tasks in the L1 have a greater potential of offering information than L2. These findings serve as one explanation of Shohamy's findings that open-ended questions in the L2 are the most difficult and MC in the L1 are the easiest.

Open-ended dual LoA Questions. Godev et al. (2002) incorporated a previously-unused study design among third-semester, intermediate students of Spanish at a university. Four groups of students were given the same reading comprehension exam with seven open-ended questions, only the language of the assessment task was separated by question and response. The language of the questions and responses were different for each of the groups and were divided as follows:
questions in English/answers in English (Group EE, n = 21), questions in Spanish/answers in Spanish (Group SS, n = 16), questions in English/answers in Spanish (Group ES, n = 20), and questions in Spanish/answers in English (Group SE, n = 15). The perceived effect of LoA on the outcomes of the exam varied from question to question. Three of the seven questions produced differences in the outcomes in a significant way. For example, question 7—*How did Martin find out the address to the hospital?*—induced a significant amount of correct responses among Group EE as compared to SS, and among Group ES as compared to Groups SS and SE. In each of these scenarios, the participants answered correctly less frequently when the question was posed in L2 (Spanish), leading to the conclusion that the L2 question led to confusion regarding what was being asked. Several in Group SS did not even answer the question, and others gave very illogical answers. Students in Group SE gave responses in English that clearly indicated that they did not understand a keyword from the question, leading to lower outcomes on this question among this group.

Question 2—*Why did Martin go to Seville?*—revealed findings that were quite different from those of Question 7. There was a significant difference among the groups such that those who answered the question in the L2 scored significantly higher based on the following comparisons: Group SS scored higher than Group SE, and Group ES scored higher than Group SE. While not significantly different, Group EE scored higher than both of the groups that answered in the L2. Analyzing the responses, it was clear that the incorrect responses stemmed from two key misunderstandings in text, knowing the definition of a key verb and understanding possession markers. The students who answered in the L2 offered more vague responses, perhaps, serving as evidence that they did not know how to respond. It was also easier for them to use the unknown verb in their answer because they did not have to produce its definition. The same applies to the grammatical concept of possession. Students could perhaps more easily copy
structures from the text in order to answer the question. Herein lies one of the potential drawbacks of open-ended questions in the L2. Also, copying and pasting, as a strategy, can lead to correct responses, but teachers while leaving teachers unsure whether the students comprehended the passage.

Statistically significant differences were also found on question 4 — *What is Rosario’s occupation?* — between Group EE and Groups SE and ES with Group EE scoring higher. In this case, it was also concluded that a key verb from the question and the text caused the students in Groups SE and ES difficulty in answering correctly. Students in Group SE that did not know the key verb were not given it in English as were their counterparts in Group ES and EE, but were required to answer the question that included that verb. Students from Group ES were given the key verb in the question in their L1, but then had to answer the question using L2 words, which means they had to know that the verb in the text was the same as the verb in the question. It seemed that only when the question was given in the L1 and when they answer this question in the L1 were they more likely to answer it correctly.

This study makes some compelling contributions to the issue of LoA in that it indicates that selecting a language for assessments is not always a clear cut choice. Three statistically significant findings were discovered among three different questions, each of which tells a different narrative. At the very least, the narrative demonstrates that the selection of LoA is more involved than a straightforward decision between L1 and L2. The study further indicates that vocabulary is a clear factor, regardless of the language used for assessment, particularly among beginning and intermediate level learners, whose vocabulary knowledge is comparatively limited.

**Written recall.** Few studies have been dedicated to LoA on written recall tasks. Lee (1986) investigated the interaction between language of recall and level of student. The
participants were instructed to read the passage and then write as much as they could recall from the L2 reading passage without referring to the passage. Lee found strong evidence that amount of recall suffered when conducted in the TL as compared to groups that answered in the L1. Those that answered in the L1 tended to write substantially more than the others. These findings were similar across the four semester levels of students.

Brantmeier (2006) conducted a study among advanced learners of Spanish at a university using written recalls to determine the effect of LoA. Initial calculations revealed that 3% of the variance in the written recalls were attributed to LoA, indicating that recall language did not matter. However, when a pre-study placement test for L1 reading ability was taken into consideration by evaluating recalls against that ability it was found that students of lower reading ability performed worse on L2 recalls than L1. A fairly sizable 28% of variance was attributed to prior L1 reading achievement, a finding that concedes even advanced learners can be affected by LoA.

Attitudes and LoA

Researchers of LoA have not yet put much emphasis on qualitative data collection. In fact, little is known about the attitudes of test takers toward LoA, specifically their preference and opinions about the language used in their foreign language (FL) reading comprehension tests. LoA researchers have advanced several ideas about the disadvantages to the test taker and to researchers in accurately measuring the trait (reading comprehension) when questions are in L1 versus L2. For example, Brantmeier (2006) and Shohamy (1984) posited that unfamiliar vocabulary words create a scenario where test takers may not understand one or more words when the question is in L2, thus leading them to choose the wrong answer even though they may have understood the passage. Brantmeier additionally mentioned that questions in the L2 encourages surface reading of passages because the learner can match words and phrases from
the test item. Nevo (1989) found that questions in L2 led to more guessing or matching of similar words and phrases. Godev et al. (2002) and others postulated that the wording and morphosyntactic structures of the questions may be cause for misunderstanding of L2 questions (Godev et al, 2002; Poh and Hock, 1979; Stroebe, 1930; Wolf, 1993). Shohamy (1984) stated that test takers may experience more anxiety when faced with questions in the L2, especially at the beginning levels, and this could potentially hinder test performance.

Though it may seem that questions in L1 would be free of criticism because they are not subject to as many linguistic issues, researchers have identified some potential challenges. Gordon and Hanauer (1995), for example, pointed out that questions in the L1 offer information to test takers that they can use in selecting the correct answer. He continued that “because open-ended tasks in the L2 are the least informative, they could be considered the most difficult to answer. Similarly, because MC tasks in L1 are highly informative, thus constituting a richer source of information for the test taker, they are potentially easiest to answer” (1995, p. 317). Shohamy also corroborated this idea by suggesting that questions in the L1 often contain clues that clarify vocabulary and overall meaning in the reading passage. In other words, questions in L1 could have repercussions on construct validity.

In connection with MC items, Nevo (1989) elicited questionnaire information about participants’ strategy use when questions were presented in both L1 and L2. Questions in L2 unveiled a significant relationship between answering correctly and incorporating contributory strategies. In contrast, questions in L1 gave no sign of such a relationship. Nevo was also able to rank strategies according to frequency of use. One of the most popular strategies in both languages was finding clues in the text, specifically in the section of the text referenced in the question, and then searching in the surrounding information. Interestingly, participants hardly ever used the strategy of ‘guessing not based on any rationale’ when questions were in L1, but
utilized it much more in L2. Additionally, participants were less likely to return to the text while answering the questions when they were posed in L2, possibly resulting from “lack of confidence, inadequate linguistic basis and incomplete understanding” (p. 208). Aside from these few differences, strategies overall transferred from L1 to L2.

Filipi (2012) also included survey data in connection with statistical measures of difficulty in order to better understand the test taker’s perspective on the difficulty of questions in L1 and L2. Those intermediate students whose L2 was French and Japanese largely preferred questions in L1. Not much additional information was gleaned from questionnaire data that would explain the reason for the preferences. Nonetheless, this type of qualitative research serves as a useful example of areas in which LoA reading research would benefit from further investigation.

Summary

Based on previous research, it was anticipated that LoA would not have a great effect on test scores among university students at the advanced level. However, LoA studies have yet to examine Russian-language texts that utilized questions empirically validated for reading at a given proficiency level. Moreover, the preferences of advanced level learners toward LoA have yet to be investigated. It is assumed, however, that as students approach native-like proficiency they will become more indifferent and possibly even prefer L2 for questions based on the naturalness of the experience.
Chapter 3: Methods

The present study explored different effects of questions in Russian on reading comprehension scores of advanced learners of Russian and their attitudes toward LoA. This chapter describes demographic and other pertinent details about those the participants of this study. It also details the instruments, procedures and data analysis used to answer the research questions regarding the effect of LoA on reading comprehension and students' attitudes toward LoA.

Participants

The participants were 64 male and female students who were enrolled in Russian 330 - Cultural History of Russia at Brigham Young University during the Fall or Winter semester of the 2013-2014 academic year. Recruiting began in the Fall of 2014 and continued into the Winter of 2014, but all participants participated in the study during the Winter semester of 2014. The average age of the participants was 21.74 (SD = 1.11). Almost all of the participants had previously lived in a Russian-speaking country for one and a half to two years serving on church missions where they gained most of their proficiency.

The Instrument

The research instruments for this study comprised a reading comprehension assessment and survey items. To measure student ability, a reading comprehension test was developed that took into consideration the advice of Godev et al. (2002) who after completing their own LoA study state:

A middle ground solution to the limitations associated to L1 or L2 testing instruments is to take advantage of both languages in such a way that the instrument may offer a set of all-L2 questions and another one of all-L1 questions, with both sets carefully designed so that they cancel out each other’s shortcomings. (p. 213)
To measure attitudes and other affective characteristics, survey questions were created. Please note that the test instrument was designed to gather more data than was needed to answer the initial research questions posed in this study. This additional data was collected in connection with and as a part of the assessment and survey for this paper and will be reported in future studies.

**Reading Comprehension Assessment.** The reading comprehension assessment consisted of two equivalent test forms (Form A and Form B) each with 20 questions that had been piloted for use on ACTFL exams and professionally arbitrated by content experts (three language experts and one proficiency testing expert) for validity. These questions were the same used in the aforementioned Clifford and Cox (2012) study and were certified appropriate for the advanced and superior proficiency scales (see Appendix A).

The reading comprehension exam incorporated a counterbalanced design (see Figure 2) such that questions on both forms were the same, but for Form A the first 10 questions were given in Russian, and the last 10 were in English and for Form B, the converse was true. An additional element of the design, the findings of which were not addressed in this report, was to give the students the same question in the other language right after they answered it in the first language. For example, the students of Group 1 who had questions 1-10 in English, were also given the same question again right after they answered in English, but this time they were to see and answer in Russian. The order of the answer options was scrambled in order to discourage the participants from simply putting the same answer when they saw the questions back to back in different languages.
The questions and passages for the reading comprehension exam were selected on the basis that they appropriately reflected the ACTFL reading proficiency scale for the advanced and superior levels. There was only one MC question per passage (n=20), therefore both the question and the passage varied in length and difficulty. Of the 20 questions and passages on the reading comprehension assessment, four were Level 2 (ILR) or Advanced (ACTFL), and the remaining 16 were for Level 3 (ILR) or Superior (ACTFL). The purpose of having four advanced level questions and passages was to offer the students some warm-up and cool-down questions that essentially sandwiched the more difficult questions. The advanced-level questions and passages were numbers 1, 10, 11, and 20, thus they were the first and last questions of each 10-question section. In addition to sandwiching the superior-level questions and passages with advanced-level questions and passages, a pyramid approach was implemented in selecting the order of the superior-level questions and passages so that they increased in difficulty up until questions 5 and 6, and 15 and 16 for each 10 question segment, and then afterwards descended in difficulty until the last questions of each 10-question section (see Figure 3). The difficulty rankings used in establishing question order were calculated by using an Item Response Theory analysis that had been previously conducted.
The questions for each passage were multiple-choice (see Figure 4). Each question contained five answer options with the 5th option always being 'I don’t know.' This option was offered to participants if they truly did not know which answer was correct, and was included as part of confidence measures. Russian faculty members at Brigham Young University, including two native Russian speakers, translated the question stems and answer options into Russian. The topics for the passages and questions fell into one of five categories. They were 1) arts and entertainment, 2) individual and society, 3) business and economics, 4) current events and politics, and 5) science and technology.
The survey instruments were designed to delve into what the examinees were thinking and feeling as they took the reading comprehension assessment. Some of the surveys are beyond the scope of this current study but were gathered for further review. These include the internal pop-up survey, confidence and anxiety items as well as some post-survey items.

Internal pop-up survey. As participants were able to see the same question twice, one after the other in both languages, this survey was designed to examine why examinees may have changed their response. If participants changed their answers upon seeing and answering the same question in the other language, an internal survey popped up onto the computer screen with survey items that were designed to elicit linguistic and other factors that led them to change their answers. For the intent of this paper, only the results of the first viewing and answering of each question were used and therefore the internal survey was also excluded.
**Internal anxiety and confidence survey.** After each reading comprehension question, the participants were also asked to answer two brief survey items that gave them the opportunity to express their level of confidence in their answer choice, and their level of anxiety while answering the question. The participants essentially answered the survey items twice for each
question because each question was viewed and answered twice, once in the L1 and once in the L2. Each of the two survey items (confidence and anxiety) was presented using a six-point Likert scale that incorporated a sliding bar as opposed to a MC button format. The questions looked as follows:

1) How confident are you in your answer choice?

Possible answer options: very unconfident – unconfident – somewhat unconfident – somewhat confident – confident – very confident

2) Indicate your level of anxiety while answering this question.

Possible answer options: very low – low – somewhat low – somewhat high – high – very high

Figure 6. Screenshot—Confidence and Anxiety Measures

Post Exam Survey. Following the 20 reading comprehension questions in both languages the participants were asked to complete a post-test survey. The purpose of the survey was to evaluate their overall attitudes toward the language of assessment in terms of difficulty,
preference, and comprehensibility of the questions. These survey questions included six-point Likert items. Each survey item contained a statement followed by the option to select ‘strongly agree’, ‘agree’, ‘somewhat agree’, ‘somewhat disagree’, ‘disagree’, and ‘strongly disagree’. After selecting the degree to which they agreed or disagreed with a given statement, students were presented with an open dialog box where they could expound on their answer choice.

Figure 7. Screenshot—Survey Items

Elements of instrument not included in this paper. At the time of gathering data for research questions treated in this paper, other information was collected for future analysis. The findings will not be discussed at this time, however to understand the procedure the research participants went through, it is important to understand the entire instrument.

Procedure

The possibility of receiving either of the two assessments was randomized by having the test administrator alternate Exam A and Exam B based on arrival time so that every other student received opposite exams. For example, the first student received Exam A and the second student
received Exam B, and the third student also received Exam A, and so on and so forth. The randomization happened naturally as the students decided on their own when to take the exam during the two-week period that it was available. The study was administered to the participants on computers in the Humanities Testing Center at Brigham Young University. The students were not given a time limit, but based on data analysis of previous research that used the same reading comprehension test items, a time estimate of 90 minutes was calculated. The average time that each participant spent on the study closely matched the initial estimation of 90 minutes.

In order to incentivize the participants to do their best on the exam, extra compensation was offered for a superior rating/score. All participants were given $20 for participating, and those that scored in the superior range were given an additional $5 dollars for a total of $25.

Data Analysis

To answer the first research question, the means of both groups were calculated on the Russian LoA test and English LoA test. Afterward a repeated measures analysis of variance (ANOVA) was utilized for evaluating test scores. To answer the second research question, the means of the post-exam survey (6-level Likert scale) on attitudes were calculated and the qualitative comments were analyzed for possible trends.
Chapter 4: Findings and Discussion

This study examined two questions relating to LoA for advanced learners of Russian. The first was concerned with the effect of LoA on reading comprehension test scores, and the second pertained to the assesses’ attitudes toward LoA on reading comprehension exams.

Research Question 1: The Effect of LoA on Scores

The descriptive statistics revealed means for each group and language that uncovered several things. First, Groups A and B which were determined by random self-selection produced different means. Group A (mean=5.28), on the whole, scored higher on the reading comprehension exam than Group B (mean=4.63) regardless of whether questions were presented in English or Russian (see Table 2 and Figure 8). Second, students scored substantially higher on items where questions were in English (mean=5.55) as compared to when they were in Russian (mean=4.37). Furthermore, the stronger group (Group A) similarly scored higher on questions in English as well as in Russian.

Table 2
Numeric Means by Group

<table>
<thead>
<tr>
<th>Lang</th>
<th>Stat. Category</th>
<th>Group A</th>
<th>Group B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>30</td>
<td>34</td>
<td>64</td>
</tr>
<tr>
<td>Eng</td>
<td>Mean</td>
<td>5.83</td>
<td>5.26</td>
<td>5.55</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>2.41</td>
<td>2.16</td>
<td>2.28</td>
</tr>
<tr>
<td></td>
<td>95%CI</td>
<td>[4.95, 6.71]</td>
<td>[4.52, 6]</td>
<td>[4.99, 6.12]</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>30</td>
<td>34</td>
<td>64</td>
</tr>
<tr>
<td>Russ</td>
<td>Mean</td>
<td>4.73</td>
<td>4.00</td>
<td>4.37</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1.84</td>
<td>2.04</td>
<td>1.96</td>
</tr>
<tr>
<td></td>
<td>95%CI</td>
<td>[4.05, 5.41]</td>
<td>[3.30, 4.70]</td>
<td>[3.88, 4.86]</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>30</td>
<td>34</td>
<td>64</td>
</tr>
<tr>
<td>Total</td>
<td>Mean</td>
<td>5.28</td>
<td>4.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>2.08</td>
<td>2.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td>95%CI</td>
<td>[4.6, 5.96]</td>
<td>[3.99, 5.27]</td>
<td></td>
</tr>
</tbody>
</table>
The means were based on a 10 question sets and therefore can be easily converted into percentages such that Group A scored 58% and 47% and Group B scored 52% and 40% on English and Russian items respectively. The total percentages then for Groups A and B were 52% and 46% respectively based on 20 questions.

In order to strengthen the significance and generalizability of the means, a mixed methods, repeated measures ANOVA was calculated. The dependent variable was the number of correct responses out of 10 for each language set. The independent variables were language in which the question was presented—English or Russian, and the group—Group A or Group B. The findings indicated that there was a significant difference in scores for both groups when questions were posed in English as opposed to Russian. Both groups performed significantly better on questions that were posed in English \[F(1, 62) = 21.47, p < .001, \text{ partial } \eta^2=.26\] indicating a large effect size (Cohen’s \(d\) .55, \(r = .26\)). In addition, estimated marginal means were compared for each group and language (see Figure 8) to determine whether either of groups

![Figure 8. Mean Ability by Group and LoA](image)
individually performed better when questions were in L2 even though, on the whole, scores were higher when questions were in L1. Subsequently, no such interaction was found and even the stronger group (Group A) attained higher scores when questions were presented in English.

![Estimated Marginal Means of LoA](image.png)

*Figure 9. Estimated Marginal Means*

**Research Questions 2: Affective Variables**

Determining the attitudes of test takers toward LoA was a product of describing statistical measures from the post-exam survey as well as finding meaning to the optional comments that followed each survey item.
Likert means. The post-exam survey initially produced a table expressing each participant's level of agreement with each survey statement. Responses were in word form (somewhat disagree, somewhat agree, etc.) and therefore warranted additional data reorganization in order to better express overall attitudes toward LoA. To do so, each Likert response was assigned a numeric value such that Strongly Disagree = 1, Disagree = 2, Somewhat Disagree = 3, Somewhat Agree = 4, Agree = 5, and Strongly Agree = 6. With each survey response converted to a number, a simple calculation of means was conducted in order to determine the average level of agreement for each survey statement. The mean response for each survey statement was regarded as substantial if it was 4 and above, or 3 and below. For example, the survey item 11 reflected a mean total of 3.46 indicating that on the whole participants' level of agreement was almost exactly between somewhat disagree and somewhat agree. In other words, they did not prefer questions in English even though they considered them easier as indicated by survey items 1-6 (see Table 3). In the graph below, the polarity of the means was reversed so as to display the means on a more uniform scale for the sake of comparison.
Table 2  
Attitudes Toward LoA

<table>
<thead>
<tr>
<th>Survey Statement</th>
<th>Mean</th>
<th>SD</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) The questions in Russian were overall more difficult</td>
<td>4.05</td>
<td>1.11</td>
<td>[3.78, 4.32]</td>
</tr>
<tr>
<td>2) The English questions were easier...</td>
<td>4.3</td>
<td>1.3</td>
<td>[3.9, 4.6]</td>
</tr>
<tr>
<td>3) I understood passages but it was difficult to transfer meaning into English questions.</td>
<td>2.76 (4.24*)</td>
<td>1.12</td>
<td>[2.49, 3.03]</td>
</tr>
<tr>
<td>4) Difficult to switch from reading in Russian to answering in English.</td>
<td>2.65 (4.35*)</td>
<td>1.3</td>
<td>[2.33, 2.79]</td>
</tr>
<tr>
<td>5) Easy to switch from reading in Russian to answering in English</td>
<td>4.54</td>
<td>1.03</td>
<td>[4.29, 4.79]</td>
</tr>
<tr>
<td>6) Didn’t know a lot of words that appeared in the question.</td>
<td>4.11</td>
<td>1.1</td>
<td>[3.84, 4.38]</td>
</tr>
<tr>
<td>7) I did not know some of the words in the Russian questions.</td>
<td>4.29</td>
<td>1.28</td>
<td>[3.97, 4.6]</td>
</tr>
<tr>
<td>8) Not knowing even one word was difficult.</td>
<td>3.52</td>
<td>1.32</td>
<td>[3.2, 3.85]</td>
</tr>
<tr>
<td>9) Even when I understood the question in Russian it was difficult.</td>
<td>3.4</td>
<td>1.1</td>
<td>[3.13, 3.67]</td>
</tr>
<tr>
<td>10) English questions were more frustrating.</td>
<td>2.4 (4.6*)</td>
<td>0.96</td>
<td>[2.16, 2.64]</td>
</tr>
<tr>
<td>11) I prefer having the questions in Russian.</td>
<td>3.46</td>
<td>1.18</td>
<td>[3.17, 3.75]</td>
</tr>
</tbody>
</table>

*The polarity of some of the means were reversed in certain cases in order to give an overall visual of how and if the means were close to one another for similar survey questions.

**Difficulty.** Survey items 1 and 2 focused on difficulty of the questions when questions were posed in L1 and L2. On the whole, students tended to view questions in Russian as more difficult than their counterparts in English. Survey items 3-9 sought to understand more specifically what elements led to the participants' opinions regarding the difficulty of LoA. Survey items 3-5 inquired about elements of questions in L2 that might make them more difficult. The means revealed that switching between the L1 and L2 as well as transferring meaning from the Russian passage into English questions were not sources of difficulty.

**Vocabulary.** Survey items 6-9 were geared toward vocabulary as a source of difficulty. Apparently, when participants did not know one word in the Russian question (Survey item 1), they, overall, did not consider that a great hindrance in answering the question (level of agreement – 3.4). However, as revealed by survey item 7 (level of agreement – 4.24), when participants did not know “some” of the words in the Russian questions, they considered those
questions more difficult, and for some students, this was a frequent occurrence (survey item 6 – 4.11). These findings corroborate what other researchers have said regarding vocabulary as a main factor involved in lower scores when questions are presented in L2 (Alderson, 2000; Bernhardt, 1995; Brantmeier, 2006; Shohamy, 1984). Further research that investigates vocabulary as a variable of LoA could expound on this topic by determining what types of words participants did not know. Presumably, not knowing words that are key to understanding the passage or at least to understanding the question being answered would have a more detrimental effect on difficulty than words of secondary centrality.

**Introduction to Interpreting Attitudes**

One of the major purposes of this study was to contribute to previous research regarding the effect of LoA on test scores among a different population. It was also a major aim to understand students’ preferences in regards to LoA. It was hoped that understanding the reason behind the language preference would accomplish the following: offer explanation for the revealed effect of LoA, begin a conversation about validity and face validity as it relates to LoA, and bring to light any other issues that had not been thought of previously.

It was anticipated that scores on MC questions would possibly be unaffected by LoA based on two ideas from previous research. One—prior LoA studies showed evidence of a diminishing effect of LoA on test scores as students became more advanced (Shohamy, 1984; Lee, 1986). In the case of Shohamy’s research, there was no significant effect of LoA on the advanced group of twelfth graders. Two—though Brantmeier’s study among advanced learners yielded evidence of a negative effect on test scores when tasks were in L2, it utilized the more difficult question type of written recall. In addition to this research, there was an assumed advantage for participants of the present study because most of them had lived in a Russian speaking country for one and half to two years. The outcome that questions in L2 were more
difficult was even more interesting in light of the finding that overall there was no preference toward one language over the other. One student’s comment even encapsulated this seemingly odd juxtaposition by stating “I didn't really have much of a preference. It was easier in English, however.” A look into the students’ preferences and their reasons for them shed light on this seeming contradiction.

**Preferring Questions in L2**

Thirty-one students responded to the optional open-ended questions for a response rate of 50%. These responses shed some light on learners’ preferences with regards to LoA. Among the major categories that arose were 1) naturalness, 2) strategies involving vocabulary, and 3) motivation.

**Naturalness.** Eighteen of the 31 comments were from those that preferred questions in L2. Three of the 18 comments made reference to questions in L2 being more natural. One of the higher performing students who preferred questions in Russian commented that “It's more natural to discuss Russian passages in Russian.” The naturalness of an L2 activity is certainly a reflection of an advanced language learner’s mindset. Research has shown that advancing language learners typically come to a point where they think more and more in the L2 and thus tasks in L2 presumably become more “natural.” By contrast, beginning learners are less likely to consider activities in the L2 as natural when they are grappling with all of the newness of learning a language. This idea also goes hand in hand with research by Upton (1997) and Upton and Thompson (2001) which showed that students rely less on L1 in reading comprehension as they progress toward advanced levels. Interestingly, this student who appreciates the naturalness of the Russian questions scored the exact same in both languages. Two other comments echoed this sentiment: “Russian answers seem a bit easier since the text was in Russian” and “it was easier for me to just answer in Russian, largely because I just had to think in one language
instead of switching between two.” The act of switching, perhaps, reflects the previous idea of the naturalness of the test instrument. In addition to being less natural, going back and forth between languages could be seen as cumbersome, thus potentially violating the terms of a good test question as defined by Shohamy. A good test, she says, is one where “the method has very little effect on the trait,” and a bad test is one where the method “has a strong effect on the trait being measured and consequently on the test takers’ scores on such tests.” If test takers view this so called code-switching during examination as unnatural and placing additional strain on them to the point of having an effect on the trait, then this act could be considered a danger to the testing instrument. Curiously, this student who did not like “switching” back and forth actually scored higher when questions were in English.

**Strategies involving vocabulary.** Several strategies surfaced as explanation for preferring questions in L2. One student noted that “[Russian questions] helped with words that I wasn't sure about in the text although it took me longer to figure it [the answer] out.” Implied in this statement is that by comparing words from the question to the text the student was able to learn words or at least better understand them. Gordon and Hanauer (1995) raised awareness of the possibility of learning from the questions themselves when he asserted that MC questions, because they offer the most information to the test taker, are also a rich source of information in helping the test taker answer questions. Brantmeier has also cautioned that questions in the L2 encourage surface reading of passages because the learner can match words and phrases from the test item. Until further research is conducted that spans across the proficiency levels, intuition suggests that matching may be more of a concern at the beginning levels. In her own study, Nevo (1989) found that L2 questions led to more guessing or matching of similar words and phrases among a group of intermediate students of French (Hebrew was the L1). This may have happened with one student in the present study who admitted that “it is easier to match
vocabulary from the question to the passage than to make good guesses using the English words.” Such a comment begs the question of whether this preference to match over guessing is a feature of the LoA or rather an area where better test items should be written. Clarity on the issue may be found in another student comment: “I feel like there wasn't a huge difference overall, at some points though it helped to understand the passage better when the questions were in Russian.” A qualifying phrase in this statement distinguishes it from the previous comment by articulating the frequency of the help rendered by questions in L2—“at some points.” If questions in L2 did not always help the test taker in understanding the text or even words from it, but only “sometimes” as another student put it, then it’s possible that more carefully constructed test items, with this very test taking strategy in mind, could mitigate the possibility of students gleaning information from the test questions themselves when presented in L2.

Motivation. One last theme came to light in regards to preferring questions in L2—motivation. As one student explained, “if it's a Russian exam, I would like to answer the questions in Russian. It gives me an incentive to want to learn more vocabulary if I am going to take an exam as a Russian would.” In this statement is a reminder of the ongoing call of researchers who have asserted that educators must take into consideration the students’ purpose for studying a foreign language when planning and implementing language building activities and exercises. Perhaps, not seriously thought of before, but the test itself could serve as a means to providing a purpose-driven and ‘meaningful’ experience for language learners. Moreover, the imperative to conduct needs analysis among students is once again reiterated, even for the sake of better testing practices.

Preferring Questions in L1

Though students on the whole scored lower when questions were posed in L2, the prior section illustrates that there were some benefits to the students in the form of naturalness,
strategies and other perceived advantages. In looking at the themes from students’ comments in regards to preferring questions in L1, two major categories emerged—strategies involving vocabulary and general difficulty of the question.

**Strategies involving vocabulary.** In regards to strategies and issues pertaining to vocabulary of the questions one student responded, “I mean, I know if I had a wider vocabulary, it (Russian questions) would not be a problem, so up to a point yes (I would prefer questions in Russian).” Another student stated that “at…times the English was good if I didn't know an important word.” This supports Brantmeier (2006) and Shohamy (1984) who criticized questions in L2 by stating that unfamiliar vocabulary words creates a scenario where test takers may not understand one or more words in the question, thus potentially leading them to choose the wrong answer even though they may have understood the passage. In cases where students did not know one or more words, particularly important ones in the question, questions in L1 could offer additional clarity, ensuring the students at the very least understood the questions. There is, however, a danger similar to what was expressed regarding questions in L2 as one student noted: “if there is a word you don't know in the passage an English question could help you figure it out.” This closely parallels the comments of those that preferred questions in L2 who incorporated a comparable strategy to learn words from the passage, and also corroborates Shohamy’s belief that questions in the L1 often contain clues that clarify vocabulary and overall meaning in the reading passage. In this case, as compared to the case of questions in L2, the student may have incorporated different metacognitive activities such as translation, and matching or guessing based on different information than if questions were in L2. The study of Godev et al. (2002) with open-ended items showed that cognates, quasi-cognates, false cognates and quasi false cognates can either aid or lead the test taker astray when questions are in English. The same could certainly be true of MC items.
**General difficulty of the questions.** Some comments spoke to the difficulty of the questions: “It was harder in Russian” and “honestly, the questions in Russian were easy, but the answers in Russian were difficult.” The latter comment may initially seem like commentary on the construction of the MC items from the exam or even MC items in general. However, this seems unlikely in light of the fact that no such comment was made in reference to the questions in English which had been formerly arbitrated by experts and rated at comparable difficulty to the questions in Russian. This leaves aspects that are specific to LoA as a basis for establishing personal preference.
Chapter 5: Discussion and Conclusion

Most of the prior LoA studies have measured reading comprehension test scores in both L1 and L2 in order to see if scores were higher or lower in a particular language. The consensus has been moving toward having questions in L1 on the basis that scores were higher when questions were in L1. Now, based on the survey findings of the present study, several possible answers to the question of “why” test scores are higher in L1. Possible answers to the “why” question open up a new conversation concerning validity and assessments in L1 and L2. Previously, conclusions about LoA were largely based on which language led to higher scores. A study that only compares in L1 and L2 does “not bring us closer to knowing which method best taps the reading comprehension trait, or which method best assesses the comprehension of texts (Shohamy, 1984).” While scores veritably reveal the difficulty to the test taker, participant comments of the present study relayed that validity of the assessments may be the more valuable question.

Does L2 Have a Lowering Effect or Does L1 Have an Elevating Effect?

In evaluating students’ comments about questions in L1 and L2, two different categories of preference were discovered. They are 1) aspects of LoA that when combined with format specific characteristics of MC questions gave participants a perceived advantage or disadvantage, and 2) aspects of LoA that may have given participants an advantage, but were not obviously related to the test method (i.e. motivation and general difficulty). The potential threats to validity lie within the first category of preference. As made apparent by the survey, one strategy that was incorporated by both those who preferred questions in L1 and L2 was that of using the information in the question, including vocabulary, to answer the questions. Some students found questions in Russian “helped with words” in the question, while others seemed to appreciate questions in English when they “didn’t know an important word.” Even those that preferred such
a strategy admitted that it was only advantageous at times. In either case, the test taker was using
the characteristics of the test method to better understand the reading passage and thus ultimately
answer the questions. It is also plausible that some students may have used the questions in L1 as
well as questions in L2 at various times throughout the research assessment to have an advantage
in answering the questions. Because of the nature of the optional and exploratory survey the
extent to which L2 had a lowering effect or L1 an elevating effects on scores is undetermined.
The most likely scenario, however, is that participants’ final scores may have been a product of
both lowering and elevating factors, and that scores reflect a composite effect of LoA and not a
solitary linear effect. If this is the case, questions in L1 appear to be more advantageous, or
advantageous more often thus leading to overall higher scores.

**Implications for Researchers**

Multiple-choice questions have received criticism irrespective of LoA based on the MC
method having an impact on the trait. Now, with L1 and L2 as overlaying variables, the impact
of the method on the trait should be considered in connection with the implications regarding
validity. It is very possible that students were incorporating problem solving strategies instead of
demonstrating reading comprehension (Rupp et al., 2006). In the spirit of decreasing extraneous
factors that lead to problem solving strategies instead of comprehension strategies, it may serve
researchers well to investigate which LoA instigates the desired strategies. It appears that L1
may have a more invalidating effect on the test simply because test takers are more inclined to
utilize the assessment, or method, to display mastery of the trait. In order to make a case for the
most valid assessment, the trait, ideally, should be as free from the method’s interference as
possible.

**Face validity.** The student from above who commented that questions in English were
overall easier, also followed up by stating: “I don't know if you're going for easy though.” The
implication of this statement is that the student has some idea of the concept of validity and that difficulty alone may not be the only relevant factor in determining it. It is unclear whether the student believes questions in the L1 to be more valid because they are easier, or questions in L2 to be more valid for other reasons. Nonetheless, such attitudes are a part of the information that ultimately constitute an unresearched aspect of LoA—face validity. Face validity, by definition, includes a subjective view of the validity of an assessment, or in other words, the test taker’s perspective of whether the assessment is a fair measure. In the case of the present paper, however, only preferences were statistically described, accompanied by comments that uncovered potential questions about validity. In order to truly understand face validity in relation to LoA, a more comprehensive assessment needs to be conducted, one where more detailed information about preference is investigated. As it stands now, students were overall indifferent toward LoA but on an individual level a wide spectrum of preferences was noted. Within those preferences were many different rationale, some of which demonstrated that students’ perception of an exam’s validity and their preferences do not always correlate. Furthermore, some students preferred questions in the L2 but actually scored better in English. Forthcoming research could expound more adequately upon these findings.

**Limitations and Future Research**

In addition to the previously mentioned areas where supplemental research is needed, several other areas merit further investigation. Future research should make certain to focus more than in the past on test takers rather than on test scores, and elicit more qualitative information that will expound upon the categorical themes discovered in the exploratory survey data of the present study. For example, deeper investigation into question-related variables such as vocabulary, strategies, motivation, and naturalness is needed. These areas could be examined from several different vantage points—preference, difficulty and validity. The present study only
began the conversation about preference, therefore further exploration of it should be carried out in order to clarify issues related to face validity. Furthermore, the findings of the present study reveal that preference and difficulty regarding LoA do not necessarily correlate. Research on difficulty among the advanced levels would benefit from research on the main contributing factors to difficulty in each language. Though preference and difficulty inform the discussion of face validity, additional measures that are specific to face validity could yield additional insights.

In addition to implications for face validity, the survey unveiled potential areas of research in regards to construct validity. Multiple-choice items, independent of the L2 research, have received criticism for various reasons. Multiple-choice questions on reading comprehension exams contain more information and thus offer the test taker more information to use in order to answer the question (Godev et al., 2002). The information contained in the questions in the present study invited certain participants to implement a strategy of comparing the question information with the passage information in order to learn more information. From the survey comments it is apparent that the strategy of using questions in L1 as well as L2 to learn information was utilized to some degree. Whether that information actually helped in answering the questions correctly was undetermined. It is also not clear to what extent other students used information from the questions to better understand the passages and thus ultimately be more equipped to answer the questions. Understanding more fully the implications of choosing a language for assessment in light of the potential impact on construct validity could shape the decision of which language to use for MC reading comprehension exams at the advanced level. Validity implications as they relate to LoA is a useful topic to pursue in future research.

Aside from themes that arose in the qualitative data, other factors could explain the effect of LoA on test scores. Shohamy hypothesized that anxiety may make L2 questions harder for beginning learners. Considering that scores were lower even among advanced learners could
justify research of anxiety with questions in L1 and L2. “Intimidating” was the word used by one student to describe questions in L2 and also possibly lends support for such future research. Other affective variables such as confidence and self-efficacy, along with their relationship with LoA may prove useful areas of research.

Another aspect of LoA research that needs subsequent expanding is the languages used for test instruments. As illustrated above, few languages have been tested in LoA research. Of the seven major research studies found on LoA, four were in EFL settings and three were in FL settings where Spanish was the L2. As FL research suggests, each foreign language has a unique interaction and relationship with L1 and therefore interpreting LoA research should be considered in this light. Moreover, the nature of LoA research may be such that outcomes among other language pairs may lead to substantially different results than those previously found.

There were several limitations of the present study. First, the small number of participants (n=64) limits the generalizability of the findings and our confidence in the data. The exploratory nature of the qualitative research, however, achieved its purpose of identifying themes and shedding light on previously under researched topics. Second, the study would have greatly benefitted from further dividing participants into groups of high and low ability based on prior proficiency measures in order to better interpret the results. In cases where some students preferred L1 and some students preferred L2 for various reasons, it would have been illuminating to know if there was a correlation between previously determined proficiency level and scores on the reading comprehension exam and preference for either LoA. Relevantly, Brantmeier (2006) initially found no effect of LoA on written recall test scores until she divided participants by actual proficiency level and not only by year in school. Third, it is possible that some participants of the study were not fit for superior level questions. Participants were invited to the study based on enrollment in an advanced course in Russian cultural history. Moreover, a previously
assumed advantage for participants was the opportunity most them had to live in a Russian speaking country for one and half to two years. In the end, however, the overall scores on the reading comprehension exam were unexpectedly low, suggesting that the MC items may have been above most students’ proficiency level (overall mean = 49.35%). Having utilized test items that had been empirically validated for the superior level may be evidence that most participants were not actually superior level readers, but perhaps advanced or lower. In this case, subdividing like Brantmeier may not have even been the proper choice.

Conclusion

Reading comprehension test scores were higher when questions were presented in L1 among advanced learners of Russian. The qualitative data revealed that understanding the reasons behind the higher scores is complex and that students were overall indifferent toward LoA. The fact that scores and preference did not completely align led to a more individual analysis of survey comments which revealed that there was actually a wide spectrum of opinions about LoA. Because survey comments were optional, the data therefrom serves as a basis for better understanding the effect of LoA and for determining categorical themes for that effect.

Students who preferred questions in L2 used descriptors such as natural and more congruent with the text. Others cited strategies such as the ability to match vocabulary from the question to the text, or simply the ability to use the question to better understand the passage. Motivation also surfaced as an unexpected aspect of preference that is actually not related to linguistic traits of LoA. Some students offered their own perspectives on the validity of assessment in L2 by asserting that Russian is a better measure of overall proficiency, thus opening a conversation about face validity in relation to LoA.

Certain students who preferred questions in L1 attributed unfavorable aspects of questions in L2 as reasons for preferring L1. Difficulty of the questions in Russian was offered
as explanation for preferring L1. Academic success was even cited as the reason for preferring L1, alluding to greater ease of questions in L1. Other cases offered more specificity, in which vocabulary appeared to have a debilitating effect on certain students, thus leading them to not completely understand the question. Take away the lack of understanding vocabulary in the question, and it appears some students would have preferred questions in L2. Vocabulary may very well be one of the major factors, dictating follow up research on LoA. Similar to the advantage of questions in L2, the ability to use questions in L1 in order to better understand the passage turned out helpful to some test takers. Understanding and comparing the interaction between both L1 and L2, and students’ ability to learn more about the passage from MC questions would be beneficial to FL educators in making decisions about LoA.

In conclusion, determining the effect of LoA on MC reading comprehension exams is a function of comparing means, and more particularly of understanding the elements that contribute to or deter from that effect. The survey data from the present study revealed facets of LoA that will need to be explored in more detail in subsequent studies. Test writers can incorporate the findings of this study by being conscious of the multitude of facets involved in selecting a language for assessing FL reading comprehension and thereby make efforts to create more valid assessments.
References


## Appendix A

Summary of Reading Passage Details

<table>
<thead>
<tr>
<th>Question #</th>
<th>General Topic</th>
<th>Specific Topic</th>
<th># of words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Arts / Entertainment</td>
<td>Arts News</td>
<td>47</td>
</tr>
<tr>
<td>2</td>
<td>Individual /Society</td>
<td>Russian Language</td>
<td>144</td>
</tr>
<tr>
<td>3</td>
<td>Current Politics</td>
<td>Agriculture</td>
<td>212</td>
</tr>
<tr>
<td>4</td>
<td>Individual /Society</td>
<td>Law</td>
<td>265</td>
</tr>
<tr>
<td>5</td>
<td>Business / Economics</td>
<td>Social and Economic Policy</td>
<td>297</td>
</tr>
<tr>
<td>6</td>
<td>Current Politics</td>
<td>Russian Politics</td>
<td>320</td>
</tr>
<tr>
<td>7</td>
<td>Individual /Society</td>
<td>Families</td>
<td>297</td>
</tr>
<tr>
<td>8</td>
<td>Arts / Entertainment</td>
<td>Literature</td>
<td>232</td>
</tr>
<tr>
<td>9</td>
<td>Arts / Entertainment</td>
<td>Film</td>
<td>196</td>
</tr>
<tr>
<td>10</td>
<td>Business / Economics</td>
<td>News</td>
<td>103</td>
</tr>
<tr>
<td>11</td>
<td>Arts / Entertainment</td>
<td>Culture</td>
<td>95</td>
</tr>
<tr>
<td>12</td>
<td>Business / Economics</td>
<td>Energy Policy and Issues</td>
<td>174</td>
</tr>
<tr>
<td>13</td>
<td>Individual /Society</td>
<td>Traffic</td>
<td>202</td>
</tr>
<tr>
<td>14</td>
<td>Current Politics</td>
<td>Internet</td>
<td>217</td>
</tr>
<tr>
<td>15</td>
<td>Arts / Entertainment</td>
<td>Film</td>
<td>273</td>
</tr>
<tr>
<td>16</td>
<td>Current Politics</td>
<td>Stalin</td>
<td>495</td>
</tr>
<tr>
<td>17</td>
<td>Arts / Entertainment</td>
<td>Sports</td>
<td>263</td>
</tr>
<tr>
<td>18</td>
<td>Individual /Society</td>
<td>Social Issues</td>
<td>223</td>
</tr>
<tr>
<td>19</td>
<td>Arts / Entertainment</td>
<td>Film</td>
<td>176</td>
</tr>
<tr>
<td>20</td>
<td>Business / Economics</td>
<td>Economy</td>
<td>173</td>
</tr>
</tbody>
</table>