In Advanced L2 Reading Proficiency Assessments, Should the Question Language Be in the L1 or the L2?: Does It Make a Difference?

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

Abstract When investigating foreign language (FL) proficiency in reading in higher education, one must first determine what proficient reading entails and how to operationalize it. The American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines provide a starting point in this process, but they do not provide instructions for assessing reading. Clifford and Cox (Foreign Lang Ann 46(1):45–61, 2013) define proficient reading as "the active, automatic, far-transfer process of using one's internalized language and culture expectancy system to efficiently comprehend an authentic text for the purpose for which it was written (p. 50)." According to this definition, reading is an asynchronous, written two-way interaction between author and reader, in which the reader's primary task is to comprehend the author's intent. However, since the cognitive processes involved in reading cannot be directly observed, researchers use observable tasks (e.g., answering questions, reading aloud, etc.) to make inferences about the FL learner's reading proficiency. Shohamy (Lang Test 1(2):147–170, 1984) notes that this reliance on indirect methods of assessment places a "heavy burden on the testing method and therefore may create greater variations in scores obtained as a result of these methods" (p. 149). Thus, researching how test method affects test scores is paramount to ensure that any variance in scores is due to differences in proficiency rather than choice of test method. In designing tasks to assess reading comprehension, the issue of question language (QL) arises. That is, scholars must decide whether the QL should be in the same language as the reading passage—the learners' second language (L2) or in the native language (L1) of the learner. When the QL is in the L1, it is easier to infer what the reader has understood. When the QL is in the L2, the responses are dependent on the examinees' comprehension of both the questions and the text. However, as L2 learners gain reading proficiency, they should also better be able to comprehend questions in the L2. The present study sought to fill

T. L. Cox (✉)
Center for Language Studies, Brigham Young University, Provo, UT, USA
e-mail: troyc@byu.edu

J. Bown · T. R. Bell
Department of German and Russian, Brigham Young University, Provo, UT, USA
e-mail: jennifer_bown@byu.edu; tbell@byu.edu

P. Winke, S. M. Gass (eds.), *Foreign Language Proficiency in Higher Education*,
Educational Linguistics 37, https://doi.org/10.1007/978-3-030-01006-5_7

these gaps in the research literature by examining the effect of QL on the scores of advanced readers of Russian on a criterion-referenced test of reading proficiency. Understanding the effect of QL on readers with Advanced-level proficiency will allow practitioners to make more informed decisions about design of reading assessments in general and of high-stakes, criterion-referenced tests of reading proficiency in particular.

Keywords Russian · Russian reading proficiency · Reading proficiency · Question language · Reading proficiency test · Learner perception

1 Introduction

When investigating foreign language (FL) proficiency in reading in higher education, one must first determine what proficient reading entails and how to operationalize it. The American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines provide a starting point in this process, but they do not provide instructions for assessing reading. Clifford and Cox (2013) define proficient reading as "the active, automatic, far-transfer process of using one's internalized language and culture expectancy system to efficiently comprehend an authentic text for the purpose for which it was written" (p. 50). According to this definition, reading is an asynchronous, written two-way interaction between author and reader, in which the reader's primary task is to comprehend the author's intent. However, since the cognitive processes involved in reading cannot be directly observed, researchers use observable tasks (e.g., answering questions, reading aloud, etc.) to make inferences about the FL learner's reading proficiency. Shohamy (1984) notes that this reliance on indirect methods of assessment places a "heavy burden on the testing method and therefore may create greater variations in scores obtained as a result of these methods" (p. 149). Thus, researching how test method affects test scores is paramount to ensure that any variance in scores is due to differences in proficiency rather than choice of test method.

In designing tasks to assess reading comprehension, the issue of question language (QL) arises. That is, scholars must decide whether the QL should be in the same language as the reading passage—the learners' second language (L2) or in the native language (L1) of the learner. When the QL is in the L1, it is easier to infer what the reader has understood. When the QL is in the L2, the responses are dependent on the examinees' comprehension of both the questions and the text. However, as L2 learners gain reading proficiency, they should also better be able to comprehend questions in the L2.

The relationship between QL and reading comprehension scores was first studied by Shohamy in 1984. In her large-scale study involving 655 Israeli high school students learning English as an L2, Shohamy tested the effect of multiple choice questions in both L1 and L2 and open-ended questions in both L1 and L2 on learners' scores. She found that in addition to test method (i.e., multiple-choice, written recall, etc.), QL had a significant effect on students' scores with students scoring lower on tests with questions in the L2. However, she noted that the effect on learners' scores diminished as the learners' skills increased, positing that the difference may be erased entirely for highly proficient L2 readers.

Since Shohamy's study, QL has received sparse treatment from researchers. While some research has examined the effect of QL on examinee scores on norm-referenced tests (Brantmeier, 2006; Godev, Martinez-Gibson, & Toris, 2002, Gordon & Hanauer, 1995; Lee, 1986; Poh & Hock, 1979; Shohamy, 1984), little attention has been given to the effects of QL on criterion-referenced proficiency tests. Proficiency exams are often high-stakes and summative in nature, thus it behooves researchers to understand how the QL can affect learners' scores.

The few existing studies investigating the effect of QL on test scores have primarily focused on students at the beginning levels of language learning (Godev et al., 2002; Lee, 1986; Nevo, 1989; Poh & Hock, 1979; Shohamy, 1984) largely ignoring advanced-level readers (but see Brantmeier, 2006) for whom the effect of the QL may be less significant. Furthermore, the research that has been conducted has focused on commonly-taught L2 s such as English, French, and Spanish.

The present study sought to fill these gaps in the research literature by examining the effect of QL on the scores of advanced readers of Russian on a criterion-referenced test of reading proficiency. Understanding the effect of QL on readers with Advanced-level proficiency will allow practitioners to make more informed decisions about design of reading assessments in general and of high-stakes, criterion-referenced tests of reading proficiency in particular. This study also considered students' affective reactions to the QL.

1.1 Reading Comprehension and Question Language

As a receptive skill, reading comprehension is an internal process dependent on many internal and external factors. The purpose of reading, for instance, affects students' internal processing (Linderholm & van den Broek, 2002; Lorch, Lorch, & Klusewitz, 1993). When readers read with different goals—such as for enjoyment, learning, to evaluation, etc.—they use different internal comprehension processes. In addition, studies have also shown that background knowledge affects learners' reading processes (Anderson, 1991; Brantmeier, 2005; Bügel & Buunk, 1996; Shiotsu & Weir, 2007).

Assessing reading comprehension, whether in a learner's L1 or L2, poses particular challenges. Because comprehension processes take place internally, researchers must infer reading ability through external measures. The difficulty for researchers 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, scholars have considered how different question types affect the comprehension process. In L2 testing, QL becomes another facet that can affect the test takers, their scores, and their attitudes.

Reading Proficiency In 2012, ACTFL released its most recent Reading Proficiency Guidelines which describe five major levels of proficiency that represent a geometric progression of reading skills (e.g., Distinguished, Superior, Advanced, Intermediate, and Novice) (ACTFL, 2012).[1] A study by Clifford and Cox (2013) validated these Guidelines, using a test design which aligned author purpose, reader purpose, and text characteristics. As the authors note, "the fact that a reader can get the main idea (an Intermediate-level task) of a text generated for an Advanced communication purpose does not indicate Advanced reading ability" (p. 60). Instead, at the Advanced level, learners should be able to extract details from the text. Moreover, an Advanced-level text must exhibit characteristics of the Advanced level: the vocabulary must go beyond the high-frequency vocabulary of Intermediate-level texts, and the topics should be of broader interest than those at the Intermediate level.

As Table 1 demonstrates, at the Advanced level, readers can comprehend the details as well as the main ideas of texts, and at the Superior level, readers must be able to read between the lines, determining tone and stance. The range of vocabulary required to perform at this level across a wide variety of topics is significant. As such, the vocabulary used in L2 questions, theoretically, should not pose difficulties for the Advanced- or Superior-level reader, though the vocabulary used in L2 questions quite likely might pose difficulties for Novice- and Intermediate-level learners.

Effects of Question Language on Test Scores Relatively little attention has been given to the effect of QL on reading test scores and even less has been given to the effect of QL on the test scores of advanced-level learners. Moreover, interpreting the results of the prior literature is complicated by the variety of design variables and instruments used in prior studies. See Table 2 for more detailed information on the participants and the types of questions examined. For example, some researchers have examined the effect of QL using multiple choice questions (Gordon & Hanauer, 1995; Nevo, 1989; Poh & Hock, 1979; Shohamy, 1984), while others have used open-ended questions (Godev, Martínez-Gibson, & Toris, 2002; Gordon & Hanauer, 1995; Shohamy, 1984), written recall (Brantmeier, 2006; Lee, 1986), or think aloud protocols (Gordon & Hanauer, 1995). Findings suggest that the type of question affects the difficulty of the task as much as the QL.

In spite of the differences in design, prior research on QL generally suggests that questions in the L1, especially multiple choice questions, are easier to answer than are questions in the L2 (Gordon & Hanauer, 1995; Lee, 1986; Poh & Hock, 1979; Shohamy, 1984), and open-ended questions in the L2 are the most difficult to answer (Godev et al., 2002; Gordon & Hanauer, 1995; Lee, 1986; Shohamy, 1984). In the latter case, the difficulty of responding to open-ended questions in the L2 might be attributed to issues of production, rather than comprehension.

[1] We use capital letters to refer to the ACTFL levels. Elsewhere, lowercase is used for generic labels of learners' abilities.

Table 1 Clifford's assessment criteria—Reading (2016, p. 26)

			Conditions		
ACTFL level	Reader task	Author purpose	Text type	Content	Accuracy expectations
Superior	Understand literal and figurative meanings by reading both "the lines" and "between the lines." Recognize the author's tone and unstated positions. Evaluate the adequacy of arguments made.	Evaluate situations, concepts, and conflicting ideas. Present and support arguments and/or hypotheses with both factual and abstract reasoning.	Multiple-paragraph prose on a variety of professional or abstract subjects such as found in editorials, formal papers, and professional writing.	Multiple, well-organized abstract concepts interlaced with attitudes and feelings. Social/cultural/political issues with abstract aspects and supporting facts presented as well. Most allusions and references are explained by their context.	Understand the facts; the details; and the author's opinion, tone, and attitude.
Advanced	Understand the facts and supporting details including any causal temporal and spatial relationships.	Convey structured, factual information, supporting details, and factual relationships in extended narratives and descriptions.	News reports, magazine articles, short stories, human interest features, and instructional and descriptive materials.	Concrete information about real-world phenomenon with supporting details, as well as interrelated facts about world, local, and personal events.	Grasp both the main ideas and the supporting details.
Intermediate	Understand the main idea, orient oneself by identifying the topic or main idea.	Orient by communicating one or more general ideas.	Very simple announcements, ads, and personal notes.	Information about places, times, people, etc. that are associated with everyday events, personal invitations, or general information.	Recognize the main idea and some broad, categorical distinctions.
Novice	Recognize some random items in a list or short text.	List, enumerate.	Lists, simple tables.	Sparse or random; format or external context may reveal internal relationships.	Correctly recognize some words.

Table 2 Previous QL research

Authors, Year	Level of student	L2	L1	Question types	Participants
Poh and Hock (1979)	Post High School	English	Bahasa Malay	MC	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 the university.
Shohamy (1984)	Low, Int, and High (High School)	English	Hebrew	MC and Open ended	Students in 12th grade in a high school setting in Israel (n = 655), and they were divided into proficiency groups of low, intermediate, and high.
Lee (1986)	Beg and Int (University)	Spanish	English	Written Recall	Students enrolled in four different semester level (n = 320; 80 participants per level). Spanish classes at Michigan State and University of Michigan for first and second years.
Nevo (1989)	Intermediate (High School)	French	Hebrew	MC	Students in tenth grade in High School in Israel (n = 42).
Gordon and Hanauer (1995)	10th Graders (High School)	English	Hebrew	Think-aloud protocols (MC and open ended)	Students in 10th grade (n = 28) studying English in Israel.
Godev et al., (2002)	Intermediate (University)	Spanish	English	Open ended- with different language for stem and answer	Students in third-semester (intermediate) of Spanish at a university (n = 28).
Brantmeier (2006)	Advanced (University)	Spanish	English	Written Recall	Students enrolled in an advanced-level Spanish grammar and composition course at a private university in the Midwest (n = 106).

The negative effect of QL on test scores may, however, be mitigated by higher reading proficiency. Shohamy (1984) found that the negative effects of L2 questions on test scores diminished as the students' reading skills increased. She posited that advanced L2 learners have acquired a broad enough vocabulary that testing them in the L2 does not impede their performance. Similarly, Brantmeier (2006), in a study of the effect of task language on the written recalls of 66 advanced-level learners of Spanish, found that QL accounted for only 3% of the variance in the performance of 66 advanced learners of Spanish on a written recall task. However, a sizable 28% of the variance in L2 written recall was attributed to learners with lower levels of reading proficiency, as measured by student scores on the "Romance Languages and Literatures Online Placement Exam." Scholars suggest that the L1 plays a larger role in L2 reading for novice level readers (Corder, 1978; Upton, 1997; Upton & Thompson, 2001). As readers gain proficiency in the L2, they rely less on their L1 to process texts. Thus Bernhardt (2005) asserts that, until readers reach the "highest L2 proficiency/fluency levels" (p.141), assessment should take place in the L1.

Impact of QL on Strategies and Affect Of further interest in testing language proficiency is ensuring that systematic score variance is not due to extraneous factors such as strategies or affect (e.g., confidence, motivation, etc.). In fact, a few studies suggest that the QL may affect the strategies that learners employ while reading and processing. For example, Gordon and Hanauer (1995) found that multiple choice questions offer information to the learners that they may rely on to respond to questions. Nevo (1989) noted that readers faced with multiple choice questions in the L2 were more likely to guess by attempting to match words and phrases from the text and from the questions.

The question of learners' affective responses to QL in tests of reading comprehension has largely been ignored in the research. Shohamy (1984) suggests that test items in the L2 may increase learners' anxiety levels, thus indirectly leading to lower scores, however, she did not empirically test this hypothesis. One study of learners' attitudes towards QL in *listening comprehension* may provide some preliminary insights. As part of a study to examine the difficulty of test questions in the L1 or the L2, Filipi (2012) also surveyed her participants to ascertain their attitudes towards the test questions. She found that a majority of beginning and intermediate students of French (N = 154) and Japanese (N = 194) preferred questions in the L1, generally finding the questions in the L1 to be harder. Whether this holds true in reading comprehension has yet to be seen.

This study sought to shed additional light on the issue of QL, by focusing on advanced-level learners of Russian, a less commonly taught language that has heretofore not been included in studies on QL. Not only does Russian use a non-Roman alphabet, but it is also typologically quite different from English (the L1 used in the majority of the QL studies) and thus may pose unique challenges to the L2 reader. Moreover, we sought to understand learners' affective responses to the QL. The following questions guided our research:

- What effect does QL have on reading comprehension test scores among advanced learners of Russian?
- What are the attitudes of advanced learners of Russian toward QL?

2 Methods

To explore the different effects of QL on reading comprehension scores of advanced learners of Russian and their attitudes, two instruments were created: a reading comprehension exam and an attitudinal survey. Participants were then recruited from upper division (third-year) Russian courses. A counter-balanced design was employed after which the resulting data were analyzed.

2.1 Reading Comprehension Exam

The reading comprehension exam used items that had been previously validated for ACTFL reading proficiency assessments (Clifford & Cox, 2013) in which each L2 reading passage contained a single question (in English, the L1). The instrument consisted of four Advanced-level passages in which the author's purpose was *to inform* and the readers' task was *to understand* details, and sixteen Superior passages in which the authors' purpose was *to persuade* and the readers' task was *to infer* the authors' argument. The existing multiple-choice questions were translated from English into Russian by university faculty (two native and one native speaker of English with Superior-level proficiency in Russian) in order to equalize the item difficulty.

Two forms of the test were created: one in which the examinees responded to the questions in Russian first and the other in which examinees saw the questions in English first. To control for the influence of ordering effects, the order of the questions was constant between the two test forms independent of the QL. The test consisted of two ten-question parts resulting in an exam that was 20 questions. The structure of each part started with one Advanced question followed by eight Superior and ended with another Advanced. This attempted to mimic the structure of an ACTFL Oral Proficiency Interview (OPI) in which examinees warm up and cool down with easier items and attempt the more difficult items in the middle. Once the test was created and the item ordering fixed, two test forms were created using a counterbalanced design. Form A had part 1 with the QL in Russian and part 2 with the QL in English. Form B had part 1 with the QL in English and part 2 with the QL in Russian.

2.2 Attitudinal Survey

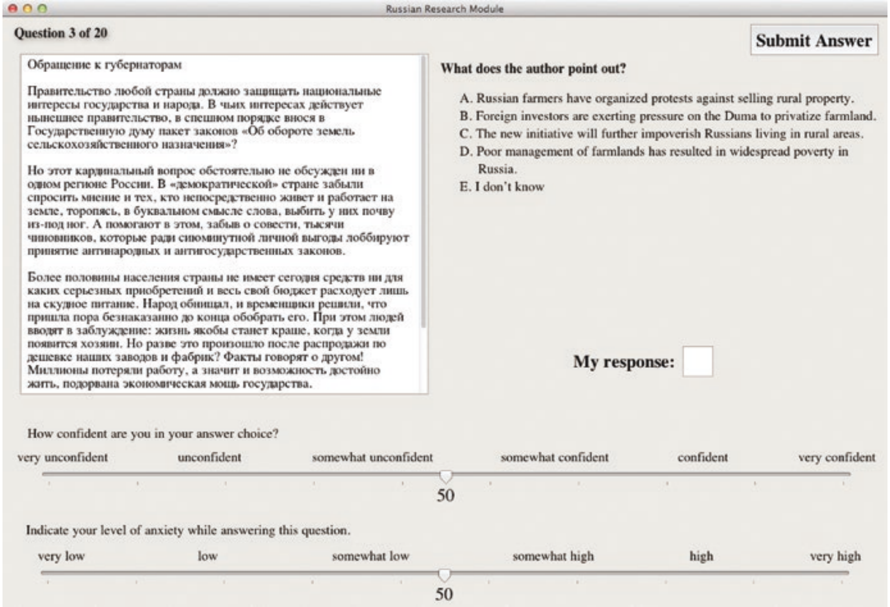
To measure participant attitudes two steps were involved. First, after every question, participants were asked to use a 100-point slider scale to rate their confidence in answering the item correctly (very unconfident to very confident) and their anxiety level with the question (very low to very high). Following the reading test, the participants were asked to complete a post-test survey. This survey was created for this study and consisted of both Likert-scale and open-ended questions. Only the open-ended questions that asked for opinions on the QL were used for this study.

2.3 Participants

The participants were 64 male (N = 51) and female (N = 13) students who were enrolled in a third year Russian language course. The average age of the participants was 21.74 (SD = 1.11) and they had all previously lived in a Russian-speaking country for an average of 22.6 months (SD = 2.92). While this group of students did not take an external proficiency test, prior OPI testing of students in this population has revealed an average speaking proficiency of Advanced-mid. Moreover, the students in the course read a great deal of material at the Advanced level. Thus, it was assumed that they were at least Advanced-level readers. All participants were given $20 for participating, and in order to incentivize the participants to do their best on the exam, an additional $5 compensation was offered for scoring in the Superior range.

2.4 Counterbalanced Design

Participants were randomly placed into two groups based on their arrival to the testing site in order to use a counterbalanced design. Counterbalancing occurs when two or more groups receive the same treatment. To avoid confounding due to ordering effects, the first group took form A with part 1 in English and part 2 in Russian, and the second group took form B with part 1 in Russian and part 2 in English. Without counterbalancing, some might argue that the differences in performance were confounded by the passages that were used instead of the QL. Counterbalanced designs use a repeated measures ANOVA with one between group variable (e.g., group membership) and one within-subject variable (e.g., QL). To answer the first research question, the aforementioned ANOVA was used on the test scores with the dependent varaiable as the test score on each part (correct out of 10 possible points), the within subjects independent variable as the QL of the part of the test and the between subjects variable as the group the participants were randomly assigned. To answer the second research question, another two repeated measures ANOVA were conducted employing the same independent variables. Confidence and anxiety were calculated by averaging the participants' responses on each post-question survey (see Fig. 1) for the two parts of the test. The open-ended responses on the post exam survey were also analyzed for possible trends.



Question 3 of 20

Обращение к губернаторам

Правительство любой страны должно защищать национальные интересы государства и народа. В чьих интересах действует нынешнее правительство, в спешном порядке внося в Государственную думу пакет законов «Об обороте земель сельскохозяйственного назначения»?

Но этот кардинальный вопрос обстоятельно не обсужден ни в одном регионе России. В «демократической» стране забыли спросить мнение и тех, кто непосредственно живет и работает на земле, торопясь, в буквальном смысле слова, выбить у них почву из-под ног. А помогают в этом, забыв о совести, тысячи чиновников, которые ради сиюминутной личной выгоды лоббируют принятие антинародных и антигосударственных законов.

Более половины населения страны не имеет сегодня средств ни для каких серьезных приобретений и весь свой бюджет расходует лишь на скудное питание. Народ обнищал, и временщики решили, что пришла пора безнаказанно до конца обобрать его. При этом людей вводят в заблуждение: жизнь якобы станет краше, когда у земли появится хозяин. Но разве это произошло после распродажи по дешевке наших заводов и фабрик? Факты говорят о другом! Миллионы потеряли работу, а значит и возможность достойно жить, подорвана экономическая мощь государства.

What does the author point out?

A. Russian farmers have organized protests against selling rural property.
B. Foreign investors are exerting pressure on the Duma to privatize farmland.
C. The new initiative will further impoverish Russians living in rural areas.
D. Poor management of farmlands has resulted in widespread poverty in Russia.
E. I don't know

How confident are you in your answer choice?

Indicate your level of anxiety while answering this question.

Fig. 1 Screenshot—passage and questions

3 Results

The participants scored substantially higher on items in which the QL was English rather than Russian (see Table 3). A repeated measures ANOVA found that while group 1 performed better than group 2 [$F(1, 62) = 451.88$, p < .001] with a large effect size (partial $\eta^2 = .88$), but more importantly, there was no interaction with the between subject variable of group [$F(1, 62) = .22$, p = .75] (see Fig. 2) and QL. Thus it did not matter whether participants saw the English questions in part 1 or part 2. Regardless of the actual reading passage, when the QL was in English participants scored higher [$F(1, 62) = 21.47$, p < .001,] with a large effect size (partial $\eta^2 = .26$).

Participants were also more confident in answering the question correctly and less anxious when the QL was English rather than Russian with a mean difference of 4.30 in their confidence level and a mean difference of 3.75 in terms of anxiety (Tables 4 and 5).

Intriguingly, the relationship between confidence and actual reading comprehension was stronger when students responded to the questions in Russian ($r = .53$, $p < .001$) as opposed to responding in English ($r = .28$, $p = .027$). In both cases, learners were overconfident in their ability. That is that they assumed they answered the question correctly when they did not. However, they were even more overconfident when the QL was English.

Attitudinal Survey A repeated measures ANOVA found that in the between subjects variable of group, there were no significant differences with either confidence

Table 3 Descriptive statistics of ability level by QL

	English			Russian		
	Group 1	Group 2	Total	Group 1	Group 2	Total
Mean	5.83	5.26	5.53	4.73	4.00	4.34
N	30	34	64	30	34	64
95% Confidence Mean (Min)	4.93	4.51	4.96	4.05	3.29	3.85
Interval for (Max)	6.73	6.02	6.10	5.42	4.71	4.84
Median	6.00	5.00	5.50	5.00	4.00	4.00
Std. Deviation	2.41	2.17	2.28	1.84	2.04	1.97
Minimum	1.00	0.00	0.00	1.00	0.00	0.00
Maximum	10.00	9.00	10.00	8.00	8.00	8.00

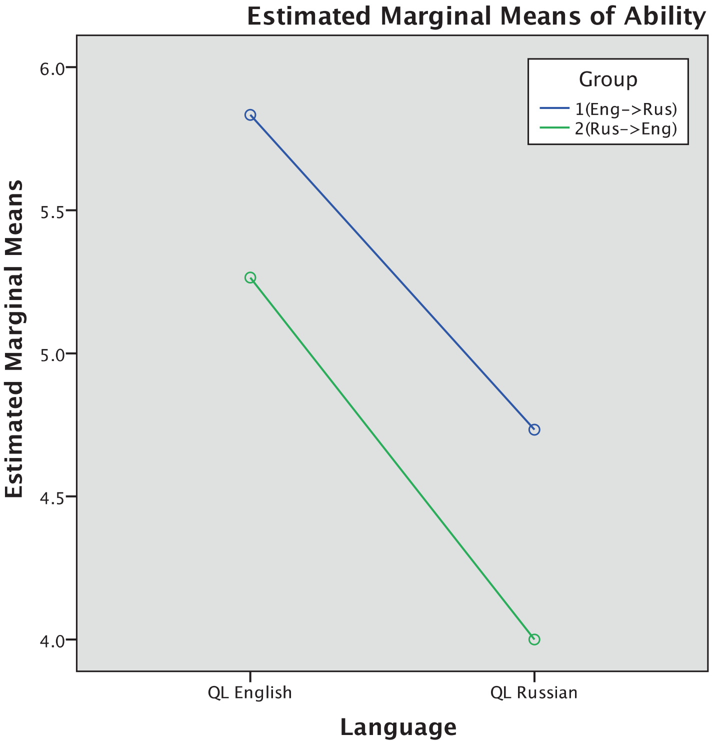


Fig. 2 Estimated marginal means of reading ability by group and QL

(F(1,62) = .002, p = .96, η^2 = .000) and anxiety (F(1,62) = .568, p = .45, η^2 = .009). We found no interaction with the between subject variable of group and QL (F(1,62) = .26, p = .61, η^2 = .004) in terms of confidence. When the QL was English, participants were more confident (the within-subject variable) in answering the question correctly (F(1, 62) = 16.33 p < .001, η^2 = .26).

Table 4 Descriptive statistics of ability level by confidence by QL

	English			Russian		
	Group 1	Group 2	Total	Group 1	Group 2	Total
Mean	61.29	60.50	60.87	55.68	56.14	55.93
N	56.17	56.06	57.60	48.93	50.95	51.85
95% Confidence Mean (Min)	66.42	64.94	64.14	62.44	61.32	60.00
Interval for (Max)	60.92	60.28	60.58	55.85	56.19	56.03
Median	59.60	58.45	59.15	51.90	54.95	52.80
Std. Deviation	13.73	12.72	13.10	18.08	14.86	16.32
Minimum	31.20	38.90	31.20	7.40	24.30	7.40
Maximum	98.70	87.30	98.70	97.30	86.70	97.30

Table 5 Descriptive statistics of ability level by anxiety by QL

	English			Russian		
	Group 1	Group 2	Total	Group 1	Group 2	Total
Mean	43.02	38.13	40.42	44.60	44.16	44.36
N	37.86	32.77	36.74	39.26	39.22	40.84
95% Confidence Mean (Min)	48.17	43.49	44.11	49.94	49.10	47.89
Interval for (Max)	47.80	44.70	47.45	49.25	48.40	49.25
Median	13.81	15.37	14.75	14.30	14.16	14.11
Std. Deviation	0.00	1.80	0.00	0.00	5.20	0.00
Minimum	60.90	60.00	60.90	63.10	68.40	68.40
Maximum	43.02	38.13	40.42	44.60	44.16	44.36

However, we did find an interaction ($F(1,62) = 8.11$, $p = .006$, $\eta^2 = .116$) between group and QL in terms of anxiety. The group that saw the English QLs first had a slightly higher level of anxiety when they subsequently encountered the Russian QL, but the group that saw the Russian QLs first reported much less anxiety when the QL switched to English. Perhaps this indicates a sense of a relief in better comprehending what was being asked of them. With both groups, though, the QL in English resulted in less anxiety ($F(1, 62) = 23.75$ $p < .001$, $\eta^2 = .28$). (see Fig. 3).

Prior to answering the open-ended questions, students were asked to respond to the following statement "I prefer having the questions in Russian." Their average on a seven-point Likert scale was 3.46 (sd = 1.18, 95%CI [3.17, 3.75]) indicating no strong preference for QL. 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 QL.

Preferring Questions in Russian Eighteen of the 31 comments were from those that preferred questions in L2. Responses of participants who preferred the questions in Russian seemed to fall into three major categories: (1) naturalness, (2) vocabulary strategies, and (3) motivation.

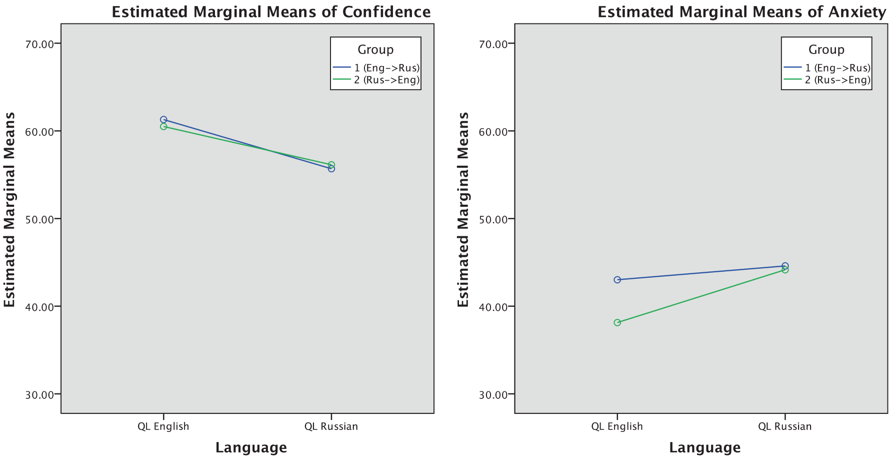


Fig. 3 Estimated marginal means of confidence and anxiety level by group and QL

Naturalness. Three of the 18 comments made reference to questions in L2 seeming more natural. One of the higher performing students who preferred questions in Russian commented that "[i]t'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 advanced language learners begin to think more and more in the L2, making tasks in the L2 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 comment echoes Upton's (1997) and Upton and Thompson's (2001) findings indicating 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, relates to the perceived naturalness of the test instrument. In addition to being less natural, switching between languages may be cumbersome, 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" (p. 147). If code-switching is perceived to place additional strain on test takers, L1 questions might invalidate the assessment 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." This statement suggests that multiple choice questions in the L2 facilitated use of problem-solving strategies. Gordon and Hanauer (1995) similarly suggested that students might learn from multiple choice questions, since they offer the most information to the test taker and serve as a rich source of information in helping the test taker answer questions. Nevo (1989) noted a similar pattern, finding that L2 questions led to more guessing or matching of similar words and phrases among a group of intermediate students of French. However, intuition suggests that matching may be more of a concern with Novice- and Intermediate-level learners, whose comprehension skills are not fully developed. Though one student in this study reported "matching" as a strategy, noting that "it is easier to match vocabulary from the question to the passage than to make good guesses using the English words," the student in question scored lower on the assessment with questions in the L2 than in the L1. If question items are well-designed, "matching" may become less viable as a test taking strategy.

Motivation. Some students who preferred the questions in Russian cited motivation as a factor in their choice. 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." This statement implies that question items in the L2 may enjoy greater face validity among Advanced-level speakers than questions in the L1. Moreover, the assessment instrument itself apparently served as a purpose-driven and meaningful experience for this language learner, further motivating the student to improve vocabulary and reading skills.

Preferring Questions in English Though students on the whole scored lower when questions were posed in L2, the prior section illustrates that questions in the L2 posed some benefits to the students in the form of naturalness, strategies, and motivation. However, some students preferred questions in their L1. In analyzing their responses, two major themes emerged—strategies involving vocabulary and general difficulty of the question.

Strategies involving vocabulary. In regards to the 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 statement supports Shohamy's (1984) and Bernhardt's (2005) concerns about testing L2 reading in the context of learners' "impoverished second language skills." (p 141). In fact, Shohamy (1984) asserts that for novice and intermediate learners, "presenting the questions in L1 may be considered more ethical, since the decision maker obtains information on the test taker's ability to understand the L2 text, without a carry-over from the language of the questions" (p. 158).

Whereas learners' inability to comprehend the question items in the L2 may have negatively affected their reading test scores, L1 question items may have offered test takers clues as to the general meaning of the text, as Shohamy (1984) posits. One student reported, "if there is a word you don't know in the passage, an English question could help you figure it out." This strategy appears similar to the L2 vocabulary matching, cited above. However, the strategies used to puzzle out new vocabulary may have been quite different when the questions were posed in the L1. Matching of L2 words may involve less comprehension than matching of L1 words to L2 words. Moreover, Godev et al. (2002) found that cognates, quasi-cognates, false cognates, and quasi-false cognates can either aid or lead the test taker astray when questions are in the L1. Overall, the student responses indicate that carefully constructed test items are important to ensure that questions do not provide too much information that can lead to strategic guessing.

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 multiple choice items from the exam or even multiple choice 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 QL as a basis for establishing personal preference.

4 Conclusion

This research study investigated the effect QL has on reading comprehension test scores among advanced learners of Russian as well as learners' preferences regarding QL. Our findings corroborate previous research indicating that questions in the L1 are easier for students. Shohamy (1984) and Bernhardt (2005) hypothesized that questions in the L2 are appropriate at the advanced levels of reading proficiency, however implicit in this assumption is that the passage, question, and examinee proficiency levels are aligned. Our data suggest that L1 questions are easier even for advanced-level learners, when responding to texts and questions that may be beyond their actual proficiency level. It may be that the QL has less of an impact when the learners' reading proficiency matches that of the intended passage and question difficulty.

We also examined students' preferences for QL in the L1 or L2. In this study, unlike in Filipi's (2012) study with lower-level learners, participants reported no strong preference for QL, in spite of the fact that questions in the L2 proved more difficult. This ambivalence towards QL suggests that questions in the L2 may enjoy greater face validity than questions in the L1 for advanced-level speakers. Face validity is defined as an individual's subjective view of the validity of an assessment, or in other words, the test taker's belief about whether or not the assessment is a fair measure of knowledge or ability (Holden, 2010). The students' lack of a strong preference may indicate their beliefs that, as advanced speakers, they *should* be able to handle questions in the L2. In fact, at least one student indicated that questions in the L2 appeared more authentic and therefore more motivating.

Even though students expressed no strong preference for QL, their confidence levels in answering correctly were generally higher when responding to questions in English than in Russian. However, that confidence was frequently misplaced with students being generally overconfident in their abilities, but tending to be more so when the QL was English. Even for Advanced-level learners, questions in the L1 may serve as a "security blanket," making them overconfident in their comprehension.

These findings are intriguing in light of Shohamy's study, suggesting that more advanced readers were "hardly affected" (p. 157) by the QL, leading Shohamy to conclude that "high and low-level" students may process L2 data differently. Drawing on Corder (1978), she suggested that learners rely more on the L1 in the beginning phases of language learning. In fact, she posits that "[i]n the beginning phases, the native language is the only linguistic system from which the learner can draw" (p. 158). Nevertheless, even advanced learners appear to feel more confident, even if over confident, in their comprehension when questions are posed in the L1.

At least one of the comments in our qualitative study discussed the difficulty associated with switching between the L1 and the L2. If, as Shohamy (1984) and Corder (1978) posit, advanced level learners draw primarily from the L2 linguistic system, switching between languages may cause psychological strain. Even if this strain does not adversely affect performance on an assessment instrument, learners may believe that code switching does. Such a belief would challenge the face validity of the instrument.

In developing criterion-referenced tests, it is important to consider QL and cut scores. Advanced- and Superior-level readers are expected to have a much broader base of vocabulary, allowing them to more easily comprehend questions in the L2. If test designers insist on presenting questions in the L1, cut scores may need to be higher to certify Advanced- and Superior-level proficiency, since it appears that questions in the L1 are easier even for advanced-level readers.

4.1 Limitations and Suggestions for Future Research

The primary limitations of this study involve a possible mismatch between the learners' reading proficiency levels and the test items. The research instrument predominately comprised Superior-level items, whereas the learners may have been at the Advanced level, or possibly below. Participants were invited to take part in the study based on their enrollment in an advanced-level course in Russian cultural history and their extended time abroad in Russian-speaking countries. In the end, however, the overall scores on the reading comprehension exam were unexpectedly low, suggesting that the multiple choice items may have been above most students' proficiency level (overall mean = 49.35%). Since the test items had been empirically validated for the Superior level, the learners' performance may be evidence that they were not actually Superior-level readers.

To better understand the effect of QL on Advanced- and Superior-level readers, researchers should first establish the proficiency level of the learners, using a criterion-referenced test and then match the instrument to the learners' level. The study would have benefitted from 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 the L2, it would have been illuminating to know find out if there was a correlation between previously determined proficiency level and scores on the reading comprehension exam and preference for either QL. Future research could also investigate the effect of QL for items below, at, and above the learners' established proficiency levels.

Additionally, future research that investigates question-related variables, such as vocabulary, strategies, motivation, and naturalness, would contribute to an understanding learner attitudes toward QL. These areas could be examined in terms of preference, difficulty, and validity. The present study only addressed QL preference and found that preference and difficulty regarding QL do not necessarily correlate. More research is needed to understand what factors contribute to the "difficulty" of a passage and item.

Student comments about the difficulty of items in the L1 or L2 only hinted at their processing and test-taking strategies. Multiple choice questions on reading comprehension exams contain a great deal of information and thus may help the test taker to answer questions correctly (Godev et al., 2002). The information contained in the questions in the present study invited participants to implement the 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 employed at least to some degree. Whether that information actually helped in answering the questions correctly was undetermined.

In the present study, we considered students' preferences for QL as well as their confidence in responding to questions posed in the L1 or the L2. However, Shohamy (1984) has hypothesized that L2 questions may cause anxiety, particularly for low-level learners. What impact anxiety might have for learners of any level remains to be determined. Other affective variables, such as confidence and self-efficacy, along with their relationship with QL may also prove to be useful areas of research.

More research is needed in this area with larger sample sizes and with a wider variety of L2 s. The QL research to date has focused on French, Spanish, and English, and while Shohamy examined QLs in Hebrew with English passages, this is the first known study to examine a less commonly taught language with a different orthographic system. As FL research suggests, each language has a unique interaction and relationship with the L1, and the effect of different writing systems such as Arabic, Chinese, Japanese, Korean, etc. should be explored. Moreover, the nature of QL research may be such that outcomes among other language pairs may lead to substantially different results than those previously found, and the interpretation of QL research should be considered in this light.

4.2 Implications for Testing and Teaching

What language should be used when testing L2 reading comprehension? Unfortunately, the answer is not entirely clear-cut and is likely dependent on the testing situation and population as well as on practical considerations. Though it may appear that the QL should be in the L2 for Advanced- and Superior-level items, there are situations in which the L1 may be preferable. For example, certain professions, particularly in government work, require a high degree of bilingual fluency. Many language professionals are required to read in the L2 and report on that reading in the L1. In such cases, requiring learners to switch languages during an assessment instrument is a valid means of assessing their ability to perform their jobs. Additionally, in the design of reading assessments for less commonly taught languages, finding testing experts with enough expertise to ensure the quality of test items may be difficult, if not impossible. In such situations, passages can be translated into a common L1, and the questions can be evaluated in that L1. On the other hand, heritage speakers of a language or students who do not speak the L1 of the dominant population may be at a disadvantage when asked to respond to questions in a third language. In order to make reading proficiency tests available to anyone regardless of L1, the instruments cannot be dependent on bilingualism.

In the end, the issue of QL remains unsolved. However this study does yield some implications for testing. First, it is important to establish the reading proficiency of learners first with criterion-referenced testing before testing items in L1 and L2. Doing so will eliminate the question of whether the test items might be too difficult for learners. Second, test designers should construct items with test tasking strategies in mind. Well-constructed items to avoid matching or giving away information and are designed with test-taking strategies in mind. And third, the type of task required to answer a test item may have an effect on outcomes. Multiple choice may be useful for large scale norm- or criterion-referenced tests, but may not always be appropriate for classroom assessment.

This study represents a first attempt to investigate the effect of QL on the scores of Advanced-level readers of Russian. Although we were unable to definitively answer the question about which language should be used in assessing reading comprehension at the Advanced and Superior levels, this study has nonetheless contributed to our understanding in this area. Advanced-level readers of Russian generally reported that questions in the L2 were more difficult to answer leading to increased anxiety and decreased confidence than were questions in the L1. However, our study finds that the level of the learners may not be as important as the alignment of the learners' proficiency level and the difficulty of the reading passages and subsequent tasks. Decisions about QL should be made deliberately, taking into consideration the level of the participants and the level of the tasks that they are expected to perform.

References

ACTFL. (2012). *ACTFL proficiency guidelines 2012*. Alexandria, VA: American Council on the Teaching of Foreign Languages. Retrieved from https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *The Modern Language Journal, 75*(4), 460–472. https://doi.org/10.1111/j.1540-4781.1991.tb05384.x

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics, 25*, 133–150. https://doi.org/10.1017/s0267190505000073

Brantmeier, C. (2005). Effects of reader's knowledge, text type, and test type on L1 and L2 reading comprehension in Spanish. *The Modern Language Journal, 89*(1), 37–53. https://doi.org/10.1111/j.0026-7902.2005.00264.x

Brantmeier, C. (2006). The effect of language of assessment and L2 reading performance on advanced readers' recall. *The Reading Matrix, 6*(1), 1–17. Retrieved from: https://pages.wustl.edu/files/pages/imce/brantmeierlanguageresearch/effects_of_language_of_assesment_0.pdf

Bügel, K., & Buunk, B. P. (1996). Sex differences in foreign language text comprehension: The role of interests and prior knowledge. *The Modern Language Journal, 80*(1), 15–31. https://doi.org/10.2307/329055

Clifford, R. (2016). A rationale for criterion-referenced proficiency testing. *Foreign Language Annals, 49*(2), 224–234. https://doi.org/10.1111/flan.12201

Clifford, R., & Cox, T. (2013). Empirical validation of reading proficiency guidelines. *Foreign Language Annals, 46*(1), 45–61. https://doi.org/10.1111/flan.12033

Corder, S. P. (1978). Language learner language. In J. Richards (Ed.), *Understanding second and foreign language learning: Issues and approaches* (pp. 71–93). Rowley, MA: Newbury House.

Filipi, A. (2012). Do questions written in the target language make foreign langauge listening comprehension tests more difficult? *Language Testing, 29*(4), 511–532. https://doi.org/10.1177/0265532212441329

Godev, C. B., Martinez-Gibson, E. A., & Toris, C. C. (2002). Foreign language reading comprehension test: L1 versus L2 in open-ended questions. *Foreign Language Annals, 35*(2), 202–221. https://doi.org/10.1177/026553229601300205

Gordon, C. M., & Hanauer, D. (1995). The interaction between task and meaning construction in EFL reading comprehension tests. *TESOL Quarterly, 29*(2), 299. https://doi.org/10.2307/3587626

Harrison, C., & Dolan, T. (1979). Reading comprehension: A psychological viewpoint. In R. Mackay, B. Barkman, & R. R. Jordan (Eds.), *Reading in a second language* (pp. 13–23). Rowley, MA: Newbury House.

Holden, R. B. (2010). Face validity. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini encyclopedia of psychology* (4th ed., pp. 637–638). Hoboken, NJ: Wiley.

Lee, J. F. (1986). On the use of the recall task to measure L2 reading comprehension. *Studies in Second Language Acquisition, 8*(2), 201. https://doi.org/10.1017/s0272263100006082

Linderholm, T., & van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology, 94*(4), 778–784. https://doi.org/10.1037/0022-0663.94.4.778

Lorch, R. F., Lorch, E. P., & Klusewitz, M. A. (1993). College students' conditional knowledge about reading. *Journal of Education al Psychology, 91*, 239–252. https://doi.org/10.1037/0022-0663.85.2.239

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing, 6*(2), 199–215. https://doi.org/10.1177/026553228900600206

Poh, T., & Hock, L. (1979). The performance of a group of Malay-medium students in an English reading comprehension test. *RELC Journal, 10*(1), 81–89. https://doi.org/10.1177/003368827901000106

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test perofmrance. *Language Testing, 24*(1), 99–128. https://doi.org/10.1177/0265532207071513

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comrehension. *Language Testing, 1*(2), 147–170. https://doi.org/10.1177/026553228400100203

Upton, T. (1997). First and second language use in reading comprehension strategies of Japanese ESL students. *Teaching English as a Second or Foreign Language – The Electronic Journal for English as a Second Language, 3*(1), 1–23. Retrieved from: http://www.tesl-ej.org/wordpress/issues/volume3/ej09/ej09a3/

Upton, T., & Thompson, L. (2001). The role of the first language in second language reading. *Studies in Second Language Acquisition, 23*, 469–495.

Troy L. Cox, PhD, is a faculty member in the Linguistics Department at Brigham Young University and serves as the Associate Director of the Center for Language Studies. He is also a certified American Council on the Teaching of Foreign Languages (ACTFL) oral proficiency trainer and has used his testing expertise as a forensic linguist and in test development projects in a number of different languages. His research interests include proficiency testing, the integration of technology with assessment, objective measurement and self-assessment.

Jennifer Bown is Associate Professor of Russian at Brigham Young University, where she serves as graduate coordinator for the Second Language Teaching program. Her research interests include literacy development in foreign languages, as well as affective issues related to language learning.

Teresa R. Bell is Associate Professor of German and Second Language Acquisition at Brigham Young University, USA. Her research focuses on effective language teaching and learning. She serves as the American Council on the Teaching of Foreign Languages (ACTFL) Program Review Coordinator for the Council for the Accreditation of Educator Preparation (CAEP).