Constructing a Russian Elicited Imitation Exam

TROY COX, JENNIFER BOWN, JACOB BURDIS

A Russian student wants to know if it is worth the expense to pay for an official ACTFL Oral Proficiency Interview (OPI). The director of a flagship program wants to measure the improvement of the oral proficiency of students returning from their in-country experience. A university department needs to provide evidence that their students are meeting learning objectives as part of accreditation. In each of these cases, a cost-effective, scalable solution to measuring oral proficiency would be helpful.

Assessing speaking ability in a foreign language has traditionally been a difficult and time-intensive task. A traditional method of measuring oral language proficiency involves role plays or interviews in the target language, which can be both time consuming and labor intensive. One of the most broadly accepted measures of oral proficiency is the American Council on the Teaching of Foreign Languages (ACTFL) OPI, which requires a thirty-minute sample of spoken speech that is then rated by at least two qualified raters. Most language-learning institutions do not have the time or resources to engage in this process with any degree of regularity, especially on a large scale. Moreover, the price of the OPI can be prohibitive for many institutions.

One intriguing approach to measuring oral proficiency is elicited imitation (EI), an approach in which test takers listen to items in the target language and repeat back exactly what they hear. The accuracy with which learners repeat the sentence has been found to highly correlate with the test takers’ oral language proficiency (Bley-Vroman and Chaudron 1994). EI instruments are much less expensive to administer and score than traditional proficiency tests; this could greatly decrease the cost of oral language assessments. Moreover, EI instruments can potentially be autorated, saving time as well as money. Several studies have reported encouraging results regarding the usefulness of EI in L2 assessment in English, French, Spanish, Dutch,
Constructing a Russian Elicited Imitation Exam

Cox, Bown, Burdis

Mandarin, German, and Japanese (Yan et al. 2015), but more research needs to be conducted with other languages. And while much of the research has focused on a broad definition of proficiency, little has been done to investigate whether EI is a suitable approach for measuring L2 proficiency based on external scales such as ACTFL, Interagency Language Roundtable (ILR) or Common European Framework of Reference (CEFR).

This study examined EI as an assessment approach, investigating how well the questions aligned with ACTFL proficiency levels and how well the test results predicted the oral language proficiency of Russian-language learners as made evident by their scores on the OPI.

Research Questions
The research questions for this article are as follows:

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

Literature Review
EI varies significantly from traditional methods in how it assesses oral language ability. Traditional methods typically simulate conversations through having the examinee interact with (1) a tester such as the OPI, (2) with two testers—one acting as an interlocutor and the other rating the conversational exchanges such as the Test of Russian as a Foreign Language (TORFL), or (3) another examinee and participate in a conversation while a tester rates the performance (TELC). To be reliably rated, high-stakes tests require a recording of the interview so the test can have multiple raters (Fulcher 2003), yet the safeguards needed to ensure the tests are scored reliably often make this kind of testing impractical (Luoma 2004). While EI does not test speaking directly, it does correlate highly with speaking proficiency and can be scored reliably by nonexperts (Son 2010) or automatic speech recognition technology (Graham et al. 2008).

EI was initially developed as a psycholinguistic tool used to measure language development but was later adapted for use as an assessment tool. The item type utilizes sentence repetition. Test takers
listen to sentences and repeat back exactly what they hear (Chaudron, Prior, and Kozok 2005). Test takers first hear short and simple sentences, but as the test continues, they are confronted with increasingly longer sentences. The increasing length and complexity of sentences facilitates testing of oral proficiency (Cox, Bown, and Burdis 2015). Test takers are graded by the accuracy with which they repeat the statements. EI is a highly intriguing approach because of the relatively small number of resources needed to facilitate it—in several languages; it has been conducted via computer and scored using speech recognition technology (see Cook, McGhee, and Lonsdale 2011; Graham et al. 2008). For a more complete review of EI and its history in the literature, consult Yan et al. (2015), Vinther (2002) and Bley-Vroman and Chaudron (1994).

**What Does EI Measure?** The purpose of this study is to investigate whether the EI test can accurately predict oral proficiency; more specifically, it seeks to determine whether items based on an established proficiency scale will predict test takers’ scores on a language proficiency assessment such as the OPI. Thus it is critical to understand what EI actually measures. Many proponents of EI tests claim that EI tests are reconstructive in nature; that is, in order to complete the task, test takers must comprehend the input by decoding the string into its constituent parts and then reconstruct it with their own interlanguage system. Although EI does not directly measure oral language proficiency, it can be used to predict and infer such skills (Bley-Vroman and Chaudron 1994; Cook, McGhee, and Lonsdale 2011; Henning 1983). The premise of the EI approach is that as sentences become more complex, the learner must make use of his or her interlanguage in order to accurately reconstruct what is heard. Therefore, those who can accurately repeat longer sentences have access to a larger bank of linguistic knowledge and competence (metaphorically speaking) and are identified as more advanced speakers of the language (Ellis 2006; Erlam 2006). Vinther (2002) describes this phenomenon as a five-step process. First, the test taker listens to the sequence of sounds that make up the prompt. Next, he or she decodes the sequence of sounds into chunks of meaningful linguistic units and stores the information in short-term memory. The test taker’s familiarity with the linguistic system (grammar, vocabulary, context, etc.) dictates how much of the information in the prompt sentence can be contained in
a single chunk. The test taker then interprets the prompt by syntactically and semantically processing the chunks from the decoding process. Finally, the test taker recalls the information and produces the sentence, utilizing his or her linguistic system to reconstruct the prompt.

Though some scholars have claimed that EI measures nothing more than the ability for rote repetition through the working memory (McDade, Simpson, and Lamb 1982), others counter that EI is more than simply rote repetition. Working memory, according to Cowan (1996), represents the portion of the memory that temporarily stores information only relevant to accomplishing a current task. Recent research indicates that working memory indeed plays an important role in EI tasks (Doughty and Long 2003); however, there is still discussion in the literature about the degree of overlap between working memory and linguistic ability as measured by EI. Erlam (2006) summarized the literature in this regard, providing three points of evidence that EI measures more than the ability to perform rote imitations. First, research has shown that working memory capacity is determined by the information in the learner’s long-term memory (Baddeley, Gathercole, and Papagno 1998). Next, Potter and Lombardi (1990) provided evidence that memory for the meaning of an utterance is retained longer than the memory for the form. Finally, Munnich, Flynn, and Martohardjono (1994) showed that sentences with incorrect grammar were corrected spontaneously during EI tasks, indicating that the learners were not merely repeating what was heard based on working memory.

Additional evidence that EI measures implicit linguistic knowledge and not just working memory was offered by Okura and Lonsdale (2012). They measured participants’ working memory abilities and their scores on an EI test in order to establish whether working memory ability had a significant impact on EI test performance. Participants took a test designed to measure working memory and an English EI test used by their institution. The correlation between EI test performance and working memory scores was insignificant ($r = .249, p = .121$). The authors reported, “The lack of significant correlations between working memory and English EI scores . . . suggests that there is more to performance on EI tests than working memory capacity” (2136).
As the research indicates, EI measures much more than just working memory. Indeed, evidence suggests that EI can provide a reasonable estimate of oral proficiency. A meta-analysis of over 76 studies that reviewed seven languages found that EI tasks are able to discriminate between speakers across proficiency levels (Yan et al. 2015). For example, Erlam (2009) conducted a study with ninety-five L2 learners of English and found a correlation of .87 between her EI instrument and the International English Language Testing System. Another study compared the use of a carefully constructed EI instrument with a more traditional speaking level achievement test (SLAT) and found a .74 correlation between the two tests (Graham et al. 2008). Cook, McGhee, and Lonsdale (2011) compared the results of EI scores and OPI scores of eighty-five L2 learners of English in order to determine the predictive ability of the EI test. They used the EI scores to compute a predicted OPI score and found a .85 correlation between the predicted OPI scores and the actual scores. Wu and Ortega (2013) used a Chinese EI instrument with L2 learners that included both advanced and beginning learners as well as heritage and foreign language learners and found EI to be useful in measuring L2 language proficiency for research purposes. Along with these findings, many others have also reported significant positive correlations between EI performance and other measures of global language assessment (Call 1985; Clay 1971; Perkins, Brutten, and Angelis 1986). However, none of the studies investigated how EI worked with learners of Russian or other Slavic languages.

**What Does Language Proficiency Measure?** The scale upon which the EI items were operationalized in this study was the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines (2012), which specify four major levels of proficiency ranging from Novice (speakers who can parrot memorized words and phrases) to Superior (speakers who can work as skilled professionals in the language). The levels are designed to be broad categorizations of language ability that indicate aligned congruence of parallel axes of Function (Ask Questions, Narrate, etc.), Content/Context (Personal, Abstract, etc.), Accuracy, and Text Type (Sentences, Extended Discourse, etc.; see figure 1). For individuals to receive a rating at a given level, they must demonstrate aligned competency among all the axes. For example, one of the functions of Superior level is Supported Opinion. While it is
quite common for beginning language learners to express likes and dislikes, the Superior function of Supported Opinion encompasses (1) expressing an opinion that could be challenged about the abstract (world of ideas), (2) using structured argumentation, (3) demonstrating a level of accuracy such that pronunciation and grammar do not distract from full comprehension of the consequences/possibilities of the stated position, and (4) using the text type of extended discourse. Novice is a prefunctional level used to categorize speakers who have memorized words and phrases but are still unable to create with the language. The functional levels include Intermediate, Advanced, Superior, and Distinguished (not currently tested). To receive a rating at one of the functional levels, an examinee must demonstrate sustained, aligned performance over a variety of topics. For a more complete description of the levels and the sublevels of low, mid, and high, refer to Breiner-Sanders et al. (2000).

Although the ACTFL Proficiency Guidelines have been in use across the country for decades, the scale is not without its critics. In her survey of the literature, Liskin-Gasparro (2003) listed several of the criticisms that the ACTFL rating scale has received over the years. First, critics have pointed out that the guidelines for the scale were based more on intuitive judgments rather than on actual data, especially with the listening and reading scales that have been accused of being nothing more than modifications of the speaking guidelines. Next, the proficiency levels have been accused of being circular in that the definition of the level is the ability of the person who is able to perform at that level. Others have criticized the validity of the rating scale because of its reliance on native speakers’ abilities as a criterion against which the performance of nonnative speakers is measured. This ideal has been clarified in subsequent definitions as the functional equivalent of a well-educated, highly articulate native speaker. Notwithstanding the criticisms, the scale has been repeatedly validated by its users and found to be a useful tool in both the private and public sector and still enjoys wide use in the United States (Liskin-Gasparro 2003; Norris and Pfeiffer 2003). We have chosen to use this scale, in spite of its flaws, because of its pervasiveness in the United States, as well as its ability to
provide both a baseline and a ceiling for language learners (unlike the TORFL).\footnote{The Test of Russian as Foreign Language requires that students register for and take a test for a particular level. If the student fails to pass the test at the required level, raters have little information about the level at which students can sustain performance. Similarly, if students pass a level, raters have no way of determining whether the students are capable of reaching an even higher level.}

<table>
<thead>
<tr>
<th>ACTFL/ILR Level</th>
<th>Function/Global Taks</th>
<th>Context/Content</th>
<th>Accuracy/Comprehensibility</th>
<th>Text Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superior/3 aka The Articulate Thinker</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced/2 aka The Informed Reporter</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediate/1 aka The Linguistic Survivor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Novice/0 aka The Preparatory Parrot</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\textbf{Figure 1. Proficiency scale with parallel axis requirements.}

\textbf{Creating a Proficiency-Based EI Test}

As noted earlier, EI involves the repetition of sentence-level texts whereas the proficiency scale focuses on text types that move from the word level to that of extended discourse. On first glance, it might seem impossible to extract criteria from the proficiency scale that would align with EI as an item type. To investigate the tenability of this prospect, we used Wilson’s (2005) constructing measures Rasch IRT validation model. This model (see figure 2) consists of four building blocks: construct (the psychological trait we are testing), items (the questions test takers respond to), item scores (the degree to which the test taker answered the
item correctly expressed in proportions) and measures (the conversion of unadjusted proportions to a linear scale that aligns person ability with item difficulty via Rasch measurement).

![Diagram](image)

Figure 2: Wilson’s (2005) Four building blocks for constructing measures.

To use Rasch IRT, two assumptions need to be met: local independence (the questions need to function independently of each other) and unidimensionality (only one trait is being measured). While the assumptions of this latent trait theory are more stringent than classical test theory, the advantage is that mathematical item difficulty parameters are person-independent and person ability estimates are item independent. This allows items and persons to be compared to each other along a common construct map. Furthermore, if the empirical item difficulties align with the proficiency levels upon which they were based (Superior item difficulties are higher than the Advanced, etc.), then there is evidence that the scale can be operationalized as hypothesized. A further discussion of these issues and how they apply to language assessments can be found in McNamara (1996); Eckes (2011); McNamara and Knoch (2012); Clifford and Cox (2013); Brown, Dewey, and Cox (2014); and Cox and Clifford (2014).

To apply this model to a proficiency-based EI test, a construct map (Wilson 2009) was designed to align EI items with the types of speakers who would be successful responding to them. A helpful metaphor is to envision a set of hurdles at different heights (see figure 3). Runners with moderate jumping ability easily clear low hurdles, but the highest hurdles can only be cleared consistently by the best jumpers. To
get the most information about a person’s jumping ability, we would want to present them with a set of hurdles that they can clear about 50 percent of the time. If runners always clear the hurdle, we will never know their upper limit. If they never clear a hurdle, we will never know their baseline ability. If we want the items to represent the proficiency scale, we must create a set of items at each functional level—a set of intermediate items, advanced items and superior items. The sublevels are determined by the quality of test takers’ performance at the base level they consistently sustain. Thus test takers who are clearing the intermediate hurdles but in essence are clipping the top edge with a toe and making the hurdle wobble would be at the low sublevel. Those who easily clear all the intermediate hurdles would be at the mid sublevel. The high sublevel is indicative of their performance at the next level. Thus Intermediate High would be those who successfully clear all the intermediate hurdles and many advanced hurdles but are not able to sustain their performance at the advanced level.

To create EI items with this model and construct map, we then evaluated past studies to determine the item characteristics that correlated with scores on the proficiency scale to create a table of specifications. If the items created from those specifications (refer to Figure 2) result in item difficulties that are aligned with the proficiency levels for which they were written, then there is evidence that we were able to successfully create a criterion-referenced, proficiency-based EI exam.

Three factors that contributed to EI item difficulty included lexical difficulty, grammatical complexity, and number of syllables. Vocabulary that is less frequently used in the language has a greater likelihood of being known at the upper proficiency levels than the lower, which has been confirmed in previous EI research studies (Graham, Millard, and McGhee 2010; Wu and Ortega 2013). Mastery of grammar that is marked or highly inflected is more reflective of the upper proficiency levels than grammar that is unmarked and common, as confirmed by Wu and Ortega (2013) and Hendrickson et al. (2008). But the greatest predictor of EI item difficulty has been found to be the item length in terms of syllables (Graham et al. 2008). Thus an item specification table was developed that aligned the proficiency scale with lexical difficulty, grammatical complexity, and length in syllables.
The suitable length in syllables for items in an EI instrument depends on the morphosyntactic features of the language. Miller (1956) has shown that the average individual is able to store about seven (plus or minus two) unrelated items at once in the working memory. Several more recent studies have suggested that four items (plus or minus one) is a better representation of the working memory’s capacity (Cowan 2001). This research suggests that the length of items in an EI test should at least be greater than the working memory capacity limit in order to measure interlanguage ability. The reported sentence length range of EIs for English-language learner studies has been between six and nineteen syllables (Graham, McGhee, and Millard 2010; Vinther 2002); however, the maximum number of syllables is higher in EI instruments that have been created for other languages. Millard and Lonsdale (2011) found that the appropriate syllable range for learners of French was between seven and twenty-five syllables. Thompson (2013) found that the syllable range for learners of Spanish was between seven and thirty-four.
syllables. Wu and Ortega (2013) found a syllable range of seven to nineteen appropriate for their research purposes for learners of Chinese.

There are several reasons to suppose that the maximum length in terms of syllables for Russian might be longer than English. First, Russian is a highly inflected language, meaning that much of the grammar consists of adding affixes to the root of the word, which can make words several syllables longer. The authors assumed that the affixes would be easier to chunk, meaning that a more proficient speaker of the language would be able to more easily chunk several syllables together because of the grammatical cohesion. For example, the one syllable –tion morpheme in English is represented by ция (tsiia) in Russian, which is two syllables in length. Many of the same words with the equivalent number of morphemes have more syllables in Russian than in English. Again, the research team assumed that chunking happens on a morphemic level rather than a syllabic level, allowing native Russians to chunk morphemes of more syllables as easily as English speakers chunk morphemes of fewer syllables (Bley-Vroman and Chaudron 1994). The piloting phase of the instrument demonstrated that, indeed, native Russians can repeat back sentences of up to twenty-six syllables with 100 percent accuracy.

Methods

Research Context. The test created and evaluated in this study was given to students learning Russian for the specific purpose of discussing religion in an intensive nine-week language-learning program located at the Missionary Training Center in Provo, Utah, and to students who had recently returned from an extensive missionary experience abroad in a Russian-speaking country.

Test Design. In this section, we will explain the procedure used to create an EI instrument for learners of Russian. We will describe how we extracted items and assigned items a difficulty score as well as the initial process used to refine the item bank.

Item Development. One way to select items for an EI instrument is through selecting items from a corpus of naturally occurring language. Because large language corpora exist for most of the world’s major languages, this approach can be advantageous since it places the burden
of item selection on these corpus tools rather than on an individual researcher. Millard and Lonsdale (2011) created an EI test for learners of French using the GigaWord corpus that was administered to ninety-four participants and found a .92 correlation between their EI instrument and the OPI in terms of its ability to distinguish between levels of language proficiency.

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

The corpus was not available for download, as it was designed only to be searched rather than browsed, so the research team developed a script to harvest the content. The script used the search parameters to look for all of the parts of speech. Figure 4 shows a screenshot of the advanced search menu on the corpus. We simply checked every part of speech on the upper left box in order for the search to produce all of the content contained in the corpus. The script was written in Python, a programming language. It systematically accessed each page of the results and scraped each of the results into a spreadsheet. The results were not already parsed into sentences. An additional script was required to detect a sentence-ending punctuation mark. When such a mark was found, the script entered the following content as a new entry. After each of the sentences was entered as a separate entry, another script was written to detect the number of vowels in each sentence. In Russian the number of vowels in a word corresponds directly to the number of syllables. The results were presented in a spreadsheet, in which one column contained each of the sentences scraped from the corpus and the other contained the number of syllables of each sentence.
The second bank of items was extracted from a religious social media website with personal stories of faith and statements similar to the language that the participants would encounter in their missionary experiences abroad. The research team copied the transcripts of thirty profiles into a document, which contained nearly 15,000 items. A script similar to the one used to parse the data from the corpus was developed to parse the language in the document into individual sentences in a spreadsheet. Another script was created to count the number of syllables in each of the sentences, and the results were entered into a spreadsheet, in which the first column contained the individual sentences and the second column contained the number of syllables to the corresponding sentence. Upon completion of the extraction procedures, the banks from both sources were formatted exactly the same, and the processes
described below were applied to each of the banks separately in a parallel manner.

**Item complexity.** After the item bank was created, the sentences were grouped in three levels according to levels 1–3 on the ACTFL scale (1 = Intermediate, 2 = Advanced, 3 = Superior). We analyzed the sentences according to three factors: sentence length in terms of syllables, grammatical complexity, and lexical difficulty. Determining the item complexity was our attempt to identify if the items aligned with their intended difficulty.

**Sentence length.** We conducted a pilot study to determine the maximum length in terms of syllables to be used in a Russian EI instrument. After creating a bank of items similar in grammatical and lexical difficulty and ranging in length from twenty-six syllables to thirty-four syllables, we recorded a native Russian speaker reading each of the items. We then tested the items with twenty participants who were native Russian speakers by playing the recording to them, and asking them to repeat the items verbatim. All twenty participants were able to repeat items that were twenty-six syllables in length with 100 percent accuracy. The average score for items of twenty-eight syllables was 93 percent, with a standard deviation of 0.06. The average score dropped to 88 percent, with a standard deviation of 0.11, for items of thirty syllables. In other words, we found that native speakers of Russian struggled repeating back sentences that were longer than thirty syllables in length. This study suggests that a Russian-language learner who is able to accurately repeat a sentence thirty syllables in length has reached a native-like performance for the instrument. As a result, we narrowed down the number of sentences in the item bank to those between nine and thirty syllables in length. All sentences from nine to fifteen syllables in length were assigned to the Intermediate level, sentences from sixteen to twenty-two syllables in length to the Advanced level, and sentences twenty-three to thirty syllables in length to the Superior level.

**Grammatical complexity.** In order to determine grammatical complexity, we used an indexed grammatical feature list created by OPI raters for Russian that outlines the grammar features that align with the functions of different proficiency levels (see appendix A). We used this list to assign the grammar features a score corresponding to the level of
difficulty from 1 to 3 ($1 = \text{Intermediate}$, $2 = \text{Advanced}$, and $3 = \text{Superior}$). Since the definition of the OPI novice level is the absence of language command indicated in the intermediate level, we did not assign a score to items at a novice level. We then analyzed each sentence and marked the presence of each of the grammatical features by entering its score in separate columns in the spreadsheet. Finally, we computed the maximum score, which we used to represent the level of difficult grammatical features in each sentence.

**Lexical difficulty.** Lemma frequency was used as the primary factor in determining lexical difficulty. Lemma frequency represents the cumulative frequency of all the word-form frequencies of words within an inflectional paradigm. For example, although a verb may have several conjugated forms, the lemma frequency couches each occurrence of the variation underneath the verb stem. This is important because we were not interested in the frequency of the variations of a word; rather, we were interested in the frequency of the word and all of its forms. We used a lemmatizer tool developed by Serge Sharoff from the University of Leeds to convert each of the word forms in the item bank to represent the lemma of the word (http://corpus.leeds.ac.uk/mocky/). Then we developed a script to search for the lemma word frequency of each of the words in the item bank, using a Russian lemma-frequency list, also created by Serge Sharoff (http://www.artint.ru/projects/frqlist/frqlist-en.php). We assigned each item a lexical difficulty score, which equaled the score for the least frequent word in the item. According to the lexical difficulty score, we assigned the sentence a level from 1 to 3 on the ACTFL scale: items containing the most frequent 3,000 words were assigned level 1; items containing the words of frequency 3,000–9,000 were assigned level 2; and items containing words with frequencies above 9,000 were assigned to level 3.

**Item selection.** We extracted twenty items for each syllable length from nine to thirty (440 items total: 220 from the Russian National Corpus and 220 from the social media website). The next step in determining which items to use in the EI instrument was to filter through the 440 items extracted and to identify those with the most discriminating power based on the proficiency scale. The first step in filtering through these items was to simply throw out items that were assigned different levels according to the ACTFL scale for syllable
length, grammar complexity, or lexical difficulty (see table 1). In other words, we only retained items in which all measures of complexity aligned with the proficiency scale. The rationale for doing so was to increase control in the EI test. If an item was intermediate in terms of syllable length but superior in terms of grammatical complexity, then it became difficult to understand why the item did or did not perform well in the test.

Table 1. Constraints of item complexity for ACTFL levels 1–3

<table>
<thead>
<tr>
<th>ACTFL Level</th>
<th>Number of Syllables</th>
<th>Grammatical Complexity</th>
<th>Lexical Frequency (Lemma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate</td>
<td>9–15</td>
<td>Command of Level 1 features</td>
<td>0–3,000</td>
</tr>
<tr>
<td>Advanced</td>
<td>16–22</td>
<td>Command of Level 2 features</td>
<td>3,000–9,000</td>
</tr>
<tr>
<td>Superior</td>
<td>23–30</td>
<td>Command of Level 3 features</td>
<td>9,000+</td>
</tr>
</tbody>
</table>

We then implemented an additional filter by running the items by at least two native Russian speakers to have them eliminate items that were confusing or did not make sense when taken out of context. We also removed items that contained collocates that would likely be chunked as an individual unit by most speakers of the language (for example, United States of America or organizational names.). Finally, we relied on the expertise of two specialists trained in rating the OPI in Russian. These individuals went through the remaining items and confirmed the score assigned to them as mentioned above (Intermediate, Advanced, and Superior) according to the OPI rating standards. We discarded all of the items for which all three scores (the grammar score mentioned above and the score of the two raters) did not agree. See appendix B for the final list of items used in this EI test.
Test-Administration Procedure
In this section, we will outline the procedure followed to test the items in the item bank to discover which items have the most discriminating power.

Participants. The participants in this study came from two groups. The first group consisted of fifty-two individuals: twenty-eight men and twenty-four women (ages 18–26) learning Russian in an intensive program preparing them for missionary experiences in Russian-speaking countries. At the time of the study, these participants had been learning Russian for four to eight weeks. The second group consisted of forty-four individuals: thirty-three men and eleven women (ages twenty-one to thirty-four) recently returned from extensive experiences abroad in Russian-speaking countries. Three of the participants were native Russian speakers.

EI test. The EI test consisted of seventy-two items divided into three subtests—three groups of twenty-four items divided according to the ACTFL levels as discussed previously. The items were recorded by a male native speaker of Russian reciting each of the items at a normal speed with distinct, authentic, but not slurred or distorted pronunciation. A browser-based administration program administered the intermediate subtest first, followed by the advanced and then the superior. The display for the test taker was very simple and included a reference of how many items had been completed and how many remained (see figure 5). The system randomly chose one of each level’s twenty-four items as a prompt for the test taker, and then the test taker repeated the prompt as accurately as possible. There was a delay of three seconds between each item, and then the system chose another item from the twenty-four items in that level and continued doing so until all twenty-four items from that level had been completed. The system then moved to the next level and repeated this process until all seventy-two items had been completed.

Administration. The EI test was administered in a computer lab with twelve computers, on which the test was preloaded prior to the session. The fifty-four students in the intensive program took the test in eight waves. The forty-four students who had recently returned from being abroad took the test in seven waves.
Scoring. To score the EI sentence, each rater was presented with a clickable, parsed-by-syllable version of the item. The rater was also presented with a button that played the original audio recording and a button that played the recording from the test taker. The rater listened to the recordings and clicked each syllable of the item that was either pronounced incorrectly or was not pronounced at all. The rater ignored syllables that were repeated, inserted, or too difficult to understand. If the learner’s utterance retained the same meaning but used slightly different wording, the rater would still only count as correct the syllables that were in the original prompt. Once satisfied that the item had been successfully rated, the rater moved on to the next item. This continued for each of the seventy-two items in the test. Since Son (2010) found that nonnative speakers can rate EI samples as reliably as native speakers, two nonnative Russian-speaking raters (who did not participate in the study) rated the EI tests. A third rater arbitrated any syllables that were not scored the same by the raters.

With this scoring method, individual syllables were the unit of measurement. Therefore, if there were more than one mistake in a single syllable, the entire syllable would be counted as incorrect. If a syllable was missing, that entire syllable was deemed incorrect. If a syllable was correct but placed out of order, then only one of the syllables would be
marked as incorrect. Each item had a proportional score of the number of correct syllables 0 to 1.0 that was then converted to a 4-point rating scale where 0 indicated a score lower than .1, 1 indicated a score between .11 and .50, 2 indicated a score between .51 and .90, and 3 indicated a score higher than .91.

**Official oral proficiency rating.** Within three days of taking the EI test, all participants (both those studying for four weeks in the intensive program and those who had returned from abroad) took ACTFL oral proficiency tests. Most participants took the computerized Oral Proficiency Interview (OPIc) in the same computer lab in which they had taken the EI test, with the exception of 11 participants who had already taken an OPI within three months of taking the EI. The OPIc is a test similar to the OPI, except that it is administered online. Instead of being interviewed by a live interlocutor, test takers are asked questions by a computer avatar, and their responses are recorded and rated afterwards. At the time of this study, the Russian OPIc was only able to assess language proficiency up to the advanced level.

**Results**
In order to answer the questions in this study, we used the Rasch IRT model to calculate the item-difficulty statistics of the seventy-two items on the EI test. Before reporting the findings for each of the research questions, we will present a diagnosis of the functionality of the rating scale followed by a reliability analysis of the test scores from the use of the scale.

**Scale diagnosis.** The diagnosis indicated that the 4-level scale mentioned above (0–3) functioned satisfactorily within the guidelines (Linacre 2002). The average measures as well as the threshold estimates for each of the categories increased monotonically in each case. For each of the categories, the threshold estimates were within the recommended 1.4 to 5 logits between each category, implying a distinction between each of the categories. Additionally, the spacing of the thresholds was regular, allowing the scale to be treated as interval data (see figure 6). An examination of the category probability distributions showed that each category functioned well. The outfit statistics for the category ranged from 0.84 to 1.30, none of which were out of the acceptable range.
Reliability analysis. The person ability estimates ranged from $-5$ to $9$ on the scale with a mean of $0.18$ (see figure 7). In the ninety-six exam results, only two of the outfit mean squares exceeded $2.0$, and the average for the set was $0.96$. The internal separation reliability between the test takers was $0.99$ with a separation strata index of $11.1$ indicating the EI reliably separated individuals into different groups based on their performance using the 4-point rating scale described above. The item ability estimates ranged from $-7$ to $6$ on the scale with a mean of $0$ (see figure 8). The item separation reliability statistic was also $0.99$ with a separate strata index of $9.92$. The separate strata indexes for both person ability estimates and item ability estimates were higher than expected, and we verified the analysis to make sure this was not an error. We attribute the strength of the strata index to the wide range of proficiency levels of the learners and to the three-level process we followed to determine item difficulty. Of the 72 items on the exam, only three had outfit mean squares exceeding $2.0$, and the average for the set was $1.05$. These findings imply that the items were reliably distinct from each other and can easily represent at least the three different difficulty levels that were intended.

Figure 6. Russian EI rating category distribution
Question 1: Alignment of intended and actual item difficulty levels. A Pearson product-moment correlation coefficient was computed to assess the relationship between the item-difficulty logit measures and the intended ACTFL level for each item. The data met the assumptions for using such a test in that the data were continuous and a scatterplot of the data affirmed a linear relationship. There was a positive correlation between the two variables, $r = .773$, $n = 72$, $p < .001$. Increases in intended ACTFL level were correlated with increases in the item difficulty logit measure. Additionally, a one-way, between-subjects ANOVA was conducted to compare the effect of intended item ACTFL level (1–3) on the item’s item-difficulty logit measure. The data passed the assumptions for using an ANOVA test in that the logit measures were normally distributed with only a slight right skew with no extreme outliers. There was a significant effect of intended item ACTFL level on item difficulty logit measure at the $p < .05$ level for each of the three levels $[F(2, 69) = 52.69, p < .001, \eta^2 = .60]$. Post hoc comparisons using the Bonferroni test found statistical differences between Intermediate (1) and Advanced (2) items (mean difference $= –2.38$ logits, a 95% CI $[-3.32, –1.41]$, and $p < .001$) and between Advanced (2) and Superior (3) items (mean difference $= –1.60$ logits, a 95% CI $[-2.55 , –0.65]$, and $p < .001$). Taken together, these results suggest that the actual difficulty levels as a whole align well with the intended item difficulty through ACTFL proficiency construct. Viewed in context of the first question of this study regarding the alignment of the actual item difficulty measures with their intended difficulty levels, these data indicate that the alignment is quite strong. However, a box plot of the data (figure 9) shows that for each level, there were some items that had item difficulty measures higher than the mean measure of the next intended ACTFL level.

As seen in figure 9, there were outliers in each group of intended difficulty. These items, their transliterations, and their translations are listed in appendix C. Further analysis indicated that the problematic items were not appropriately aligned with their intended difficulty levels. In at least one case, the particular curriculum that the learners were exposed to made an item easier than intended. This item contained the word вдохновляющим (vdokhnovliaiushchim “inspiring”), which increased the lexical difficulty score to level 3. However, this word was
familiar to this particular group of students, who were learning religious language.

Figure 7. Russian EI person ability map

**Question 2: Predictive ability of EI test for OPI scores.** We used the Rasch IRT model to calculate the person ability estimates for the ninety-six participants in the study. The person ability estimates were normally distributed, and a scatter plot showed a strong linear relationship. Passing the assumptions, a simple linear regression
analysis was conducted to find an equation to predict a subject’s OPI score based on the person ability estimate of the criterion-referenced, proficiency-based EI test developed in this study. Subjects’ OPI scores from the person ability estimate could be predicted by the following equation: \( y = 0.72x + 3.91, R^2 = 0.86, N = 96, r = 0.93. \) The scatterplot in figure 10 summarizes the results. These data viewed in context of the second question of this study indicate that the person ability measure is a strong predictor of learners’ oral proficiency, as made evident by an OPI score. These data establish this EI test as a suitable testing instrument to indicate Russian oral language proficiency.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Person - MAP - Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>x</td>
</tr>
<tr>
<td>8</td>
<td>x</td>
</tr>
<tr>
<td>7</td>
<td>x</td>
</tr>
<tr>
<td>6</td>
<td>+ 3-12</td>
</tr>
<tr>
<td>5</td>
<td>xx</td>
</tr>
<tr>
<td>4</td>
<td>xxxx</td>
</tr>
<tr>
<td>3</td>
<td>XXXX(5)</td>
</tr>
<tr>
<td>2</td>
<td>XXX</td>
</tr>
<tr>
<td>1</td>
<td>XX</td>
</tr>
<tr>
<td>0</td>
<td>XX M+M</td>
</tr>
</tbody>
</table>

Figure 8. Russian EI item difficulty map

**Discussion**

The relationship between the item difficulty measures and the intended difficulty level shows a 77 percent correlation, and an ANOVA showed
that the item difficulties of the items grouped by their intended difficulty levels were significantly different from each other, with an effect size of 0.60, becoming more difficult as the intended level increased. This indicates that the items ascend hierarchically based on the ACTFL scale. In regard to question 1 of this study (how well the intended item difficulty levels align with the actual levels), these results provide good evidence that the item-selection procedure proposed in the literature (Christensen, Hendrickson, and Lonsdale 2010; Millard and Lonsdale 2011) and employed in this test was sufficient to produce an effective, predictive EI test, and the items performed as intended.

The regression analysis of the person ability estimates and the OPI scores ($R^2 = .86$, $N = 96$, $r = .93$) showed that the scores on the EI test strongly predicted the scores that the participants received on the OPI, providing important information for the second question of this research study. While the EI test does not measure oral language proficiency, such a high correlation between the two tests suggests that we can, with an acceptable degree of confidence, infer oral language proficiency based on the scores of the EI test. As Erlam (2006) argued, there is strong evidence that EI measures an individual’s interlanguage system and not just working memory ability. We suggest that EI is able to obtain such strong predictive power precisely because EI measures this interlanguage system, which is at the root of oral language proficiency. These results are promising and add Russian to the list of languages for which EI has been found to discriminate among proficiency levels (Yan et al. 2015).

**Limitations**

Although the results of this study are encouraging, several limitations must be taken into account. While we have indicated the difference in levels for both the person ability estimates and the item difficulties, we have not shown that the person ability scores line up with the constructs of the item difficulties. For example, even though we have indicated which items are superior-level items and we have indicated which persons were superior-level persons, we have not provided evidence that these line up. Next, because of budget and scheduling constraints, the majority of participants in this study took the OPIc instead of the OPI. For the Novice and Intermediate levels, the OPIc is discriminates
among test takers’ abilities as well as the OPI (Surface, Poncheri, and Bhavsar 2008). This is not the case for the Advanced level. The upper-level test takers who took the OPIc and received an Advanced score did not receive a delineation of low, mid, or high. On the 0–9 ACTFL scale from Novice Low to Superior, those who received a score of Advanced on the OPIc received a 6, which is the equivalent of Advanced Low. Although several of these test takers may have been able to receive a score of Advanced Mid, High, or even Superior, the OPIc was not robust enough to differentiate at the higher levels. This lack of differentiating power hampered the ability of this study to discriminate among higher-level learners as well as it could differentiate among lower-level learners.

**Item_Measure**

![Boxplot of item difficulty statistics for intended difficulty](image)

*Figure 9. Boxplot of item difficulty statistics for intended difficulty*

Additionally, we admit that the process of determining the complexity level of the items was somewhat arbitrary. More research
needs to be conducted to more accurately predict the difficulty of each item prior to a test administration and to validate Millard and Lonsdale’s (2011) success with using corpus tools as the source for effective EI items.

**Figure 10. Scatterplot of person ability estimate and OPI score**

**Implications**
The results of this study suggest that EI can serve as a proxy for oral proficiency. While we do not suggest that EI should *replace* the OPI, or indeed classroom measures of speaking ability, EI instruments can serve a useful role in the Russian language curriculum, especially if time and resources would not otherwise permit speaking to be assessed. We see its value as a screening tool to quickly and efficiently determine a language learner’s speaking level and to assess incremental progress. It can be used as a pretest to place learners into appropriate courses. Because the test takes relatively little time, it can also be administered multiple times throughout a program (perhaps at years’ end) to gauge progress and provide formative assessment to learners and evidence for
departments for accreditation purposes. Furthermore while EI may serve primarily as a measure of learners’ speaking skills, the test can also provide diagnostic information about their grammatical and lexical knowledge, particularly since, unlike open-ended speaking assessments, EI does not allow learners to avoid forms that they do not know. The EI instrument could potentially be used to assess readiness for the Russian Federation’s own Test of Russian as a Foreign Language, helping learners determine which level they should register for.

Conclusions
This study provides supporting evidence for the use of EI in language testing. The fact that the results in this study for a little-researched language (Russian) align with the results for studies of other prominent languages suggests that EI is not language dependent. The extension to L2 Russian is important not simply because another language is added but because it represents a language of greater difficulty for English speaking learners than most of the languages previously investigated. Thus the positive results of this experiment lend credence to the overall robustness of EI as a technique. While this study does not attempt to identify which factors contribute to item complexity, we were able to produce strong results through controlling for sentence length, grammatical complexity, and lexical frequency. Most importantly, this research suggests that EI can be used as a proxy indicator of oral proficiency for purposes of screening and formative assessment.

Appendix A: Russian Grammar Features for Proficiency Levels

<table>
<thead>
<tr>
<th>Proficiency Level</th>
<th>Grammar Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate</td>
<td>Gender and number agreement in high-frequency words</td>
</tr>
<tr>
<td>Intermediate</td>
<td>Verb control in high-frequency verbs</td>
</tr>
<tr>
<td>Intermediate</td>
<td>Past, present, and future conjugation in high-frequency words</td>
</tr>
<tr>
<td>Intermediate</td>
<td>Adjectives and adverbs</td>
</tr>
<tr>
<td>Intermediate</td>
<td>Relative pronouns</td>
</tr>
<tr>
<td>Intermediate</td>
<td>Simple conjunctions</td>
</tr>
</tbody>
</table>
Constructing a Russian Elicited Imitation Exam
COX, BOWN, BURDIS

Intermediate Adverbial time words (then, tomorrow, in the morning)
Intermediate Ordinal numbers 1–100
Intermediate Basic modal verbs
Intermediate Impersonal constructions

Advanced Passive voice
Advanced Aspect
Advanced Reflexive
Advanced Prefixes of motion verbs
Advanced Relative clauses
Advanced Verb control
Advanced Declensions of number in all cases
Advanced Conditional
Advanced Comparative adjectives
Advanced Declension of proper nouns
Advanced Definite pronouns
Advanced Indirect speech

Superior Participle constructions
Superior Subordinate clauses of concession/compromise
Superior Diminutive/affectionate nouns and adjectives

Appendix B: Items in the Russian Elicited Imitation Test

<table>
<thead>
<tr>
<th>Item #</th>
<th>Syllables</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>У меня пять прекрасных дочерей.</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>Я знаю, что это Его Церковь.</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>Она ощутила истинный мир.</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>В своей жизни я стараюсь служить другим.</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>Я очень люблю это Евангелие.</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>Сейчас у меня есть сильная вера.</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>Я встала с колен со слезами на глазах.</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>Я никогда не была так счастлива.</td>
</tr>
<tr>
<td>9</td>
<td>13</td>
<td>Бог любит меня и слышит мои молитвы.</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>Это здорово помогать людям верить в Бога.</td>
</tr>
<tr>
<td>11</td>
<td>14</td>
<td>У меня есть разные обязанности в церкви.</td>
</tr>
</tbody>
</table>
Молитва укрепляет мою веру в Христа.
Она уже почти не болит.
Что будем покупать на рынке?
Ты сегодня ездила к Насте?
Не знаю как, но я тебя видел.
Я еще не совсем с ума сошла.
Я просто спросил, как у тебя дела.
Ты же сказала, что тебя не будет.
Я скажу тебе ответ на твои вопрос
У нас там на даче прекрасная осень.
Да в советское время такого не было.
Я очень рад, что вы все сегодня сюда пришли.
У нас был здесь один маленький случайный концерт.
В любом спорте я всегда играл под этим номером.
Я женат уже на протяжении 18 лет.
Церковь помогла мне стать более хорошим человеком.
Мне нравится работать с молодежью в нашей Церкви.
Как и у любой другой семьи, у нас есть свои трудности.
Я знаю, что мой Отец на Небесах призвал меня к этой работе.
Я стараюсь подавать пример чистой жизни и высоких нравственных норм.
Я остаюсь дома с моими четырьмя замечательными малышами!
Многие члены Церкви помогали нам своими различными способами.
Я прочитала Книгу Мормона первый раз когда училась в восьмом классе.
Я просыпаюсь каждый день с миром и надеждой благодаря моей вере.
Отказываясь от комплимента, вы отказываетесь от Божьих подарков.
Я ещё точно не знаю во сколько я поеду.

Чем ты планируешь заняться во время отпуска?

Он ждал меня у гостиницы, где я остановился.

Здесь он чувствовал себя очень спокойно и уверенно.

Давайте всё-таки вернёмся к более радостным вещам.

Какие у вас возражения против этого термина!

Мне бы хотелось сразу сделать небольшое замечание.

К сожалению сегодня более ста детей не попали в списки.

И мы работали с пяти утра до двух часов ночи следующих суток.

Родились люди, которые не знали никакого другого языка.

Если никто не возмущается, это еще не значит, что все всем довольны.

В итоге мы должны прийти к некоторым выводам и рекомендациям.

Я люблю встречать новых людей и укреплять уже существующую дружбу.

Одна из величайших драгоценностей в моей жизни - это моя сестра-близнец.

Я вижу много благословений благодаря тому, что я в Церкви с четырех лет.

Каждое утро я молюсь, прося о терпении в преодолении трудностей.

Я провела большую часть моей взрослой жизни, служа подросткам в нашем приходе.

Я очень люблю следовать вдохновляющим примерам людей, которых встречаю.

У нас двое замечательных детей, которые не дают нам особенно расслабляться.

Моя семья – самая большая радость в моей жизни и действительно благословение с Небес.
| 57 | 30 | Мы были благословлены тремя очаровательными дочками, которых мы просто обожаем. |
| 58 | 30 | Оглядываясь назад я понимаю, что люди вне церкви часто были лучше и мудрее меня. |
| 59 | 30 | Фактически, это – одна из величайших радостей жизни – непрерывно учиться и развиваться. |
| 60 | 30 | Это – простой принцип, но моление – это то, что в любое время под силу любому человеку |
| 61 | 23 | В любом сообществе людей существуют проблемы охраны правопорядка. |
| 62 | 23 | Я очень рада, что наконец-таки закончилось это долгданое лето. |
| 63 | 23 | Есть очень много детей-инвалидов, нуждающихся в приемных родителях. |
| 64 | 24 | Это стало для меня самым потрясающим и непростым занятием в жизни. |
| 65 | 27 | Отмечу, что за последние пять лет увеличилось число часто болеющих школьников. |
| 66 | 26 | Пожалуйста припомните на президентских выборах, за кого вы отдали свой голос. |
| 67 | 27 | Папа, будучи рыбаком, стал бригадиром, когда образовался колхоз в тридцатом году. |
| 68 | 29 | Мы рады приветствовать вас сегодня на нашем празднике посвященном дню посёлка Белогорка. |
| 69 | 29 | Если сейчас у вас это мнение поменялось то, за кого бы вы сейчас проголосовали. |
| 70 | 29 | В это мгновение слышу, какой-то вопль и только потом понимаю, что это мой собственный крик. |
| 71 | 30 | Есть ли среди вас смельчаки, которые не побоятся совершить со мной в такое путешествие. |
| 72 | 30 | Защита поддерживает заявленное ходатайство о допросе указанного свидетеля.
Appendix C: Item Outliers

Item 04
- В своей жизни я стараюсь служить другим.
- In my life, I try to serve others.

Item 07
- Я встала с колен со слезами на глазах.
- I stood from my knees with tears in my eyes.

Item 32
- Я остаюсь дома с моими четырьмя замечательными малышами.
- I stay home with my four wonderful boys.

Item 45
- Мы работали с пяти утра до двух часов ночи следующих суток.
- We worked from five in the morning until two in the morning the next day.

Item 54
- Я очень люблю следовать вдохновляющим примерам людей, которых встречаю.
- I really like to follow the inspiring example of the people that I meet.

Item 62
- Я очень рад что наконец-таки закончилось это ужасное лето.
- I am very glad that finally this terrible summer has ended.

References


