Introduction

Various second language (L2) proficiency measures are used in second language acquisition (SLA) research, ranging from self-report ratings (see Gertken et al., this volume) to standardized tests (see reviews by Thomas, 1994, 2006; Tremblay, 2011; and also discussion by Hulstijn, 2011, 2012). Researchers’ purposes for measuring L2 proficiency are diverse, and often proficiency itself is just a secondary variable in SLA research programs that investigate some other central variable construct (e.g., development of grammatical subsystems, effects of interactional feedback, learners’ vocabulary size). In this chapter we describe the development of an elicited imitation (EI) test used to measure French L2 oral proficiency. This test is the newest addition in a series of EIs that are already available in five different L2s: Chinese, English, German, Japanese, and Spanish (Ortega et al., 1999; Zhou & Wu, 2009). We argue that EI offers a useful tool for systemic yet practical assessment of L2 proficiency for a variety of SLA research purposes. Among the benefits of this particular

* This work was funded by the Economic and Social Research Council, UK (RES-062-23-2996). We are extremely grateful to our colleagues on the LANG-SNAP project for their valuable contribution to this research: Rosamond Mitchell, Laurence Richard, and Patricia Romero de Mills. We would also like to thank two anonymous reviewers for their helpful suggestions.
proficiency measure are that it is quick to administer, and that, with parallel versions in multiple languages, it allows for crosslinguistic SLA comparisons and accumulation of interpretable findings across a variety of L2s other than English.

**Background**

**Problems in Defining and Measuring Proficiency in SLA Studies**

Attempts at defining the construct of language proficiency have been made for many decades now (e.g., Oller, 1976). In a series of recent articles, Hulstijn (2011, 2012) has proposed definitions of first language (L1) and second language (L2) proficiency based on the distinction between two kinds of language ability, which he calls basic (BLC) and higher (HLC) language cognition. While BLC is restricted to the processing of oral language (listening and speaking) in utterances containing high-frequency lexical, grammatical, phonotactic, and prosodic elements, HLC is unrestricted in these respects. That is, HLC includes the processing of written language (reading and writing, i.e., involving literacy skills) in sentences which may contain low-frequency lexical or grammatical elements (2012: 429). In other words, BLC is what is shared relatively uniformly among adult native speakers, regardless of background characteristics. In contrast, HLC is where native speakers show variation due to differences in educational level, occupation, free-time activities, and so on. Applying these constructs to non-native speakers, Hulstijn claims that L2 learners can also acquire HLC just as native speakers can (e.g., through education), but “it remains an open question to what extent postpuberty L2 learners can fully acquire BLC in their L2” (2011: 242; see also Leclercq & Edmonds, this volume, for a presentation of Hulstijn’s definition of language proficiency). With this in mind, it seems that when SLA researchers’ interests are in measuring general proficiency – what we consider BLC –, then the focus should be on the assessment of oral language using tests that are literacy-independent and that focus on
operations involving high-frequency elements during the integrative use of oral language, while recruiting in real time all interrelated basic linguistic levels of lexicon, grammar, phonotactics, and prosody.

Furthermore, SLA researchers’ interests and purposes for measuring proficiency are often quite different from those in the field of language testing, where many language tests originate (see Zoghlami, this volume, for a review of two such standardized tests). In SLA, proficiency is more often included as a secondary variable, or not included at all, in many research designs. Yet, some scholars have claimed that proficiency is a major lurking variable and should be included in all of our studies (Norris, 2010). Norris and Ortega (2012) single out several research-related purposes for the measurement and reporting of L2 proficiency, including “to justify the sampling of participants into a study or to assign participants to distinct groups” and to “aid readers of research when deciding the extent to which findings can be generalized to other samples and populations” (2012: 580).

Surveys of proficiency assessment methods used in SLA research by Thomas (1994, 2006) and Tremblay (2011) have continued to demonstrate the need to establish higher proficiency measurement standards and more consistency in assessing and reporting proficiency for research, rather than educational assessment purposes. For example, in Tremblay’s (2011) survey of studies published from 2000-2008 in three refereed journals (Studies in Second Language Acquisition, Second Language Research, and Journal of French Language Studies), only 36.8% of published reports included an independent measure of learners’ proficiency. By far the most popular way of estimating proficiency was using classroom level or years of instruction, a result also found in Thomas’s (2006) synthesis of L2 studies published in Applied Linguistics, Language Learning, Second Language Research, and Studies in Second Language Acquisition, and confirmed by Callies et al., this volume, in their chapter on the use of learner corpora to test L2 proficiency.
In French SLA studies, in particular, the problem is noteworthy. Tremblay (2011) reported that out of her total corpus of 144 SLA studies, 25 focused on L2 French but only 2 of those included an independent measure of learners’ proficiency (they used oral interview and accent ratings, respectively). Other studies of L2 French appearing in different journals and a few recent doctoral dissertations have made use of additional independent proficiency measures, such as (1) accuracy on a written personal narrative (e.g., Bardovi-Harlig & Bergström, 1996), (2) the C-test or cloze test (e.g., Daller et al., 2003; McManus, 2011; Tremblay, 2011), (3) a grammaticality judgement with correction (Ayoun, 2004), and (4) X-lex (a measure of vocabulary size, e.g., Rogers, 2009). Many of these proficiency estimates are literacy-dependent (e.g., C-test, cloze). In and of itself this is not necessarily a weakness, depending on the purposes of the research and provided that the populations investigated are literate. However, the heavy literacy requirement of these proficiency assessment tools means that they allow for the intervention of explicitly learned (declarative) knowledge more than non-literacy dependent (i.e., oral) assessments would, as the latter more often tap into implicit acquired knowledge (Ellis, 2005). In terms of Hulstijn’s (2011, 2012) distinction, in other words, literacy-dependent assessment tools may be inadvertently measuring forms of HLC while providing questionable evidence of learners’ capacities for BLC in the L2, precisely the dimension that Hulstijn considers most relevant for the study of SLA issues. One type of assessment that is not literacy-dependent, and is arguably a measure of BLC that taps more implicit language competencies, is elicited imitation.

**Elicited Imitation in SLA Research**

EI is a technique that requires participants to listen to a stimulus and then repeat it as exactly as possible. Most of the time the repetition is done orally but examples of written imitation also exist (see Vinther, 2002, for discussion). EI has a long tradition of use in the fields of first language acquisition (e.g., Lust et al., 1996; Slobin & Welch, 1968 [1973]) and
bilingualism (Radloff, 1991), and it has also been used in SLA at least since the 1980s (e.g., Hameyer, 1980; Savignon, 1982). Typically, numerous stimuli are included, ranging in number of syllables and featuring a variety of grammatical structures. The theoretical rationale behind EI as a measure of language capacity is that learners can only accurately imitate sentences they have comprehended and parsed through their developing grammars (Bley-Vroman & Chaudron, 1994).

The use of EI is not without controversy. One main argument against EI relates to the issue of ‘parroting,’ that is, whether the participant has in fact comprehended the stimulus or has just imitated a string of sounds (Vinther, 2002). Consideration of the role of working memory is important in this debate. If the stimulus is short enough, it could be stored in working memory and repeated without actual comprehension. In contrast, a stimulus that is long enough to exceed working memory capacity would be only accurately repeated if the learner has grammatically parsed and decoded the message and formed a mental representation of it. What length exceeds an individual’s working memory capacity, of course, is contingent on the proficiency level of the given language user. Attention to the length of the stimuli, therefore, is one way that researchers have tried to get around this issue, and the EI stimuli typically range in length (measured in syllables, words, or characters) and are presented for repetition in order of increasing length, so as to offer appropriate levels of difficulty across a wide range of proficiencies. Indeed, sentence length has proven to be a strong predictor of item difficulty in EI performance. Graham et al. (2010) reported that sentence length accounted for 73% of the variance in item difficulty in their EI test, compared to lexical frequency, which explained only 8%. Similar results were also found in Hendrickson et al. (2010) and Ortega et al. (2002).

Within SLA, a recent burgeoning of interest in using EI is related to the need to measure implicit language knowledge (Erlam, 2006; Verhagen, 2011), although the
measurement of oral proficiency through EI tests has also been of interest since earlier years (Naiman, 1974). Recently, a group of researchers at Brigham Young University has investigated the potential of EI as an adaptive language proficiency test (Christenson et al., 2010; Hendrickson et al., 2010) and one that could allow for automatic scoring (Cook et al., 2011; Graham et al., 2008). The particular line of EI research that fuelled the present study has focused on the use of EI in crosslinguistic studies. It began when Ortega et al. (1999) developed parallel forms of the same EI test in four languages – English, German, Japanese, and Spanish – in order to investigate the relationship between syntactic complexity measures and L2 proficiency across these four foreign languages. They compared EI scores to Simulated Oral Proficiency Interview ratings, TOEFL scores, and self-assessments and found that the EI data yielded high reliability, good discrimination, and concurrent validity when used with the four university-level samples of foreign language learners in that initial study. The Spanish EI test was then employed in dissertation studies by Ortega (2000) and by Bowden (2007) with similar success. Most recently, a parallel L2 Chinese EI test was created and pilot-tested by Zhou and Wu (2009). It was also found to exhibit good reliability and validity when subsequently used with two different large samples of foreign and heritage language learners in their respective doctoral dissertations (Wu, 2011; Zhou, 2011).

**Elicited Imitation in L2 French Research**

While the popularity of EI is increasing in SLA, its use in French L2 research is scarce. To the best of our knowledge, three such studies have employed EI to date. Burger and Chrétien (2001) used an EI task to measure changes in oral production after two semesters of a French content-based class. Their test consisted of 14 sentences, which together formed a coherent text. Each sentence ranged from 12-15 syllables and was scored globally as either 0 or 1 for exact repetition. In contrast, the EI used in Erlam and Loewen (2010) included both grammatical and ungrammatical sentences targeting primarily noun-
adjective agreement, the focus of a recast treatment. This EI also differed in that after hearing each statement, learners had to give their opinion about the statement before beginning the repetition (see also Erlam, 2006). Differences in French language competence among L1 French and L1 English children from two different age groups (grades 3 and 5) was the focus of an early study by Markman et al. (1975), who administered a 44-item EI test including both grammatical and ungrammatical sentences. Their results showed clear differences for both the L1 and the age grouping variables, with the French group performing significantly better than the English group and the grade-5 group outperforming the grade-3 group.

The three studies just reviewed are suggestive of the potential of an EI to measure L2 French proficiency, be it to gauge longitudinal proficiency gains resulting from a particular curricular or instructional intervention (Burger & Chrétien, 2001; Erlam & Loewen, 2010) or to ascertain global grammatical competence differences resulting from home language background or age (Markman et al., 1975). Interestingly, all three research teams designed their French EI tasks in ways that departed from the traditional EI design involving straightforward repetition of grammatical sentences ordered in increasing syllable length and containing a variety of structures and forms. Whether and how the different design formats featured across studies may affect the reliability and validity of any EI test (for example, including both grammatical and ungrammatical sentences for repetition, or inserting a comprehension question between hearing the stimulus and issuing the repetition) is currently unknown and in need of future research. Nonetheless, given the growing crosslinguistic evidence in support of the EI test developed by Ortega et al. (1999, 2002), we thought it worthwhile to continue in that tradition and adapt that instrument for L2 French. One of the main benefits of adding a comparable French version is that it will allow for crosslinguistic SLA comparisons and accumulation of interpretable findings across L2s.

Consequently, the first two authors set out to develop the parallel EI test for L2 French
with the purpose of employing it as a measure of L2 proficiency in the context of a larger funded investigation into the linguistic benefits of study abroad for UK university students of L2 French (see McManus et al., this volume, for a second study that resulted from this same project). In order to determine whether the French EI test was a reasonable, useful measure of proficiency, the following main research question was posed in the present study:

(1) Will learners’ scores on the French EI exhibit any meaningful relationship to other indices of language proficiency collected for the larger study, namely: (a) the lexical diversity they demonstrate in productive oral and written tasks, (b) their vocabulary knowledge as measured by a vocabulary test, (c) their speech rate on an oral retell of a picture-based narrative, and (d) university end-of-year marks?

Methodology

Participants

The participants were 29 French-degree students who were all recruited from the same curricular level: They had just finished their second year of university in Britain. Their L1 backgrounds ranged and included 25 native English speakers, 2 heritage French speakers (English + French), 1 native Spanish speaker, and 1 native Finnish speaker. A background questionnaire was administered to gather information about participants’ age, length of time studying French, age of first exposure to French, institutional level, other languages studied, and how often they use French outside the classroom. Information about participants’ end-of-year marks was also collected. The participants’ mean age was 21 (range 20-24), and the sample included 26 females and 3 males, which is representative of a similar gender imbalance in the French language courses and study abroad program at this university. Mean length of French study was 11 years (range 9-15 years), and mean age of first exposure was 9.5 years old. More than half of the participants (55%, 16) reported studying an additional language at the university (5 Italian, 5 German, 5 Spanish, 1 Chinese). A group of 10 native French speakers who were taking part in an ERASMUS university exchange program at the
same English university were also recruited to take the EI as an L1 baseline.

**The EI Test and Other Instruments**

The elicited imitation instrument included 30 test sentences ranging from 7 to 19 syllables (see Appendix A), with the sentence stimuli presented in order from lowest to highest number of syllables. A native French speaker created all the sentences, using the English sentences from Ortega et al. (1999) as a model. A second French speaker checked them for syllable length and naturalness. As with the other EI L2 versions, a variety of grammatical structures were targeted, frequent vocabulary was used, and all sentences were grammatical. A native French speaker was digitally recorded reading the sentences at a normal rate, and the free program Audacity\(^1\) was used to make the test audio. The phrase “repeat as much as you can” was reiterated several times during the recorded instructions as a way to encourage learners to attempt every test sentence. Two additional recording features were implemented following Bowden (2007): (a) a 2-second pause was inserted after each target sentence and before the cue – a 0.5 second beep – signalling when the repetition should begin; (b) the length of response time was based on the time it took the native speaker to speak the sentence plus extra time depending on the number of syllables. The resulting final EI test takes 9 minutes, 15 seconds to complete, and this administration time includes two minutes of instructions and practice sentences given in English (also following those used in Bowden, 2007).\(^2\) Given the short overall administration time, it is unlikely that fatigue played any factor in examinee performance on early or late appearing items.

In addition to the EI, as part of the larger study design, participants completed a range of language assessments on the same day. These included a general oral interview, an oral retelling of a picture-based narrative, a written argumentative essay, and a vocabulary recognition task. The oral interview was conducted in French and focused on questions relating to learners’ reasons for studying languages and expectations for their upcoming year
abroad in France. The oral retelling was based on a picture story about a young girl and her cat (Langley, 2000; see McManus, 2011). Learners were given time to preview all the pictures before starting their retell in French. The written argumentative essay was computer-based and timed; learners were given a prompt on gay rights and had three minutes to plan their response before being given 15 minutes to write approximately 200 words. Vocabulary recognition was measured via the Swansea Levels “X-Lex” test (Meara & Milton, 2003). This is a yes/no test where learners see a word and have to decide whether they recognize it as a real French word or not. They see 120 words total; 100 are real words and 20 false words, and the real words come from different frequency bands.

Procedure

All participants signed an informed consent document before the start of data collection. The order in which the data collection tasks (e.g., EI test, narrative, etc.) were administrated was randomised for each participant. The total time to complete all assessments took approximately 1 hour. Learners were welcome to take breaks between tasks although no breaks were requested. During data collection a member of the research team was seated with each participant the whole time. The EI audio was presented via a laptop and learners’ responses were digitally recorded. Participants were instructed not to pause the audio at any time, and the researcher ensured these procedures were followed.

Analysis of EI and Other Data

Learners’ repetitions on the EI were scored based on a five-point scoring rubric (0-4) developed by Ortega et al. (1999) and used also in all previous studies that have employed the EI parallel versions in the five L2s. The maximum score possible for the test is 120 (30 x 4). The scoring rubric is presented in Table 1 with illustrations taken from the present data, and a complete version is provided in Appendix B.

Two raters (the second author and a native French speaker) coded half the EI data
together and disagreements were discussed until both coders agreed on the score. They then proceeded to code the other half of the data independently. The inter-rater reliability for this part was 94% exact agreement, and the disagreements were solved once again through discussion. Following this procedure, each individual test took between 10-15 minutes to score.

The oral interview data were transcribed according to CHAT conventions (CHILDES, MacWhinney, 2000) and the written argumentative essay data were converted into CHAT as well. Both sets of data were analysed for lexical diversity using $D$ (an index developed by Malvern & Richards, 2000, that accounts for text length while estimating lexical diversity for an individual), as calculated via the CLAN program (MacWhinney, 2000). The X-lex program (Meara & Milton, 2003) automatically computes a learner’s final score at the end of the test. Two scores are given by the program: the raw score and the adjusted score. The adjusted score will be lower if any false words were erroneously identified as words by the testee; this was the score used in the current study. The picture-based narrative was used to estimate learners’ rate of speech, operationalized as the number of pruned syllables produced per minute (see Lennon, 1990). Using pruned syllables means that any repetitions, false starts, or L1 use were removed prior to analysis.

Table 1. Elicited Imitation scoring Rubric with French Examples

<table>
<thead>
<tr>
<th>Score</th>
<th>Description</th>
<th>Item</th>
<th>Repetition</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Perfect repetition</td>
<td>Item 9. <em>Il prend une longue</em></td>
<td><em>Il prend les longues</em></td>
</tr>
<tr>
<td>3</td>
<td>Accurate content</td>
<td>Item 9. <em>Il prend une longue</em></td>
<td><em>Il prend les longues</em></td>
</tr>
<tr>
<td></td>
<td>repetition with some</td>
<td><em>douche tous les matins.</em></td>
<td><em>douches tous les matins</em></td>
</tr>
<tr>
<td></td>
<td>(un-) grammatical</td>
<td>(“He has a long shower”</td>
<td>(“He has long showers”</td>
</tr>
<tr>
<td></td>
<td>changes</td>
<td>every morning”)</td>
<td>every morning”)</td>
</tr>
</tbody>
</table>
Elicited imitation as a measure of oral proficiency in L2 French (2014)
Kevin McManus (kmcmanus@psu.edu)

<table>
<thead>
<tr>
<th></th>
<th>Changes in content or changes in form that affect content</th>
<th>Item 26. Pourriez-vous s’il vous plaît me passer le livre qui est sur la table?</th>
<th>Pouvez-vous me passer le livre qui est sur la table?</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Item 21. “Pourriez-vous s’il vous plaît me passer le livre qui est sur la table?”</td>
<td>(“Would you pass me the book that is on the table please?”)</td>
<td>(“Can you pass me the book that is on the table?”)</td>
</tr>
<tr>
<td>1</td>
<td>Item 6. “Je ne sais pas s’il sait très bien conduire.”</td>
<td>(“I don’t know if he can drive that well”)</td>
<td>(“I don’t know…. to drive”)</td>
</tr>
<tr>
<td>0</td>
<td>Item 3. “Les rues sont larges dans cette ville.”</td>
<td>(“The streets are wide in this city”)</td>
<td>(“blah blah blah in the city”)</td>
</tr>
</tbody>
</table>

*Note.* Changes made by the nonnative speaker are underlined.

**Results**

In light of the overall aim, which was to test the reliability as well as the validity of this new French EI test following Ortega et al. (1999), we first address the main research question via a Pearson product-moment correlation analysis that examines the relationship among learners’ scores on the EI, $D$ values for the oral interview and writing, X-lex score, rate of speech, and university end-of-year marks. We would expect strong and positive correlations among some of these variables – namely the speaking-related measures – if the EI test yields scores that can be used as a shortcut for the measurement of the participants’ French L2 speaking proficiency in the larger study. Similarly, we would expect lower
relationships with the non-speaking measures (i.e., if the construct being measured is primarily speaking proficiency). We then augment the main findings by inspecting the likely sources of difficulty that can account for the EI repetition patterns observed in the data. Finally, we report on the internal consistency of the EI test scores we obtained with this sample and compare it to the reliabilities reported for the parallel EI tests developed for L2s besides French.

**Relationship between EI scores and Other Variables**

The distribution of scores from the 29 French as a foreign language university students is displayed in Figure 1. The mean was 62.90 (out of 120 maximum possible score) with a standard deviation of 17.97. The range was 36-97. Since 4 of our participants were of an L1 background other than English only, we inspected their scores separately in order to ascertain whether their EI performance might have been in some ways different from that of the L1 English participants. The participant who received the highest score (97) was one of the heritage French speakers. The native Spanish speaker received the second highest score (96). The other French heritage speaker received a score of 74, and the Finnish native speaker received a score of 77. The mean of the native English speaking group \((n = 25)\) was 59.2 (\(SD = 15.97\), range 36-85), which is slightly lower but statistically not significantly different from the mean for the full sample. The mean for the native French speakers \((n = 10)\) was 118.40 with a standard deviation of 1.65 (range: 115-120). Thus, the performance by the L1 baseline is clearly at ceiling, as expected.
To test whether there was a relationship between the EI scores and other external criterion measures, a Pearson product-moment correlation was conducted using the EI scores and five additional variables: university end-of-year marks, the X-Lex adjusted score, $D$ as measured on both the oral interview and the written argumentative essay, and speech rate on the picture-based narrative. As shown in Table 2, statistically significant and relatively large positive correlations were found between the EI scores and end-of-year marks ($r = .78$), the EI scores and $D$ in the oral interview data ($r = .62$), and the EI scores and speech rate on the oral picture-based narrative ($r = .67$).

**Table 2. Correlation Table**

<table>
<thead>
<tr>
<th>Measures</th>
<th>EI</th>
<th>Marks</th>
<th>X-lex</th>
<th>D Wrt</th>
<th>D Spk</th>
<th>Speech Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>EI</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marks</td>
<td>0.78**</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X-lex</td>
<td>0.12</td>
<td>0.13</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D Wrt</td>
<td>0.32</td>
<td>0.25</td>
<td>-0.03</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D Spk</td>
<td>0.62**</td>
<td>0.47</td>
<td>-0.09</td>
<td>0.03</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>
Speech Rate 0.67** 0.43 -0.29 0.11 0.66** 1.00

Note. EI = Elicited imitation scores; X-lex = adjusted scores for Meara and Milton’s (2003) test; D Wrt = lexical diversity in written essays; D Spk = lexical diversity in oral interviews; Speech Rate = number of pruned syllables per minute. **p < .01

Sources of Item Difficulty

An item analysis was also conducted to investigate item difficulty, expressed as the mean score for each item across participants. The results are shown in Figure 2. As a reminder, each item was scored using a range from 0-4, with perfect repetition receiving a score of 4. A wide range of item difficulties was found, as reflected in the lowest and highest mean scores of 1.07 (item 29) to 3.97 (item 1). The most difficult items appear at the top of the graph (see Appendix A to read the full test sentence). The longest sentences (as counted by syllables) are the items closest in test sequence number to 30. With this in mind, the results suggest that syllable count is not the sole predictor of item difficulty.

Factors other than length seem to have appropriately taxed participants’ proficiency-dependent ability to repeat the items and must be taken into account. For example, items 27 and 29 both have 18 syllables and thus both must be considered “long” items. Yet, the difference in their mean scores is quite large (2.76 and 1.07 respectively). In this case, considerable differences in syntactic complexity can be found: Item 27 is monoclausal (whereas item 29 contains 2 clauses) and contains fewer morphemes (9 vs. 14 morphemes for item 29). Thus, the source of difficulty for item 29 may originate in morpho-syntactic complexity rather than syllable length. Additionally, other aspects such as the interaction between phonology and syntax, as well as prosody and register may also play a role. For example, item 18, which is the second most difficult item in Figure 2 (with a mean score of 1.14), elicited a 0 score from approximately one third (10/29) of the participants. Of those learners who did attempt it, most scored 1 due, in part, to meaning changes and, in part, failure to repeat the whole stimulus. A change in meaning occurred because learners
incorrectly overgeneralized the negator ‘ne’; the stimulus *elle ne commande que ...* (“she only orders”) was repeated as *elle ne commande pas* (“she does not order”). Whilst both *ne V pas* and *ne V que* share distributional properties of negation, only *ne V pas* is a negative construction (Hawkins & Towell, 2010). Item 5 also proved challenging (mean score of 1.79) because of the initial question form *Qu’as-tu dit* (“What did you say”), which learners tended to repeat as *Qu’est-ce que tu dis* (“What do you say”), a form that is arguably more familiar to instructed learners. None of the learners received a perfect repetition score of 4 on this item, in contrast to the native speakers who all scored 4 on this item and all of the other items just discussed (27, 29, and 18).

In sum, the qualitative inspection of sources of difficulty in the EI French test provides evidence for precisely the type of differentiation that stimuli in an EI test are expected to produce, as a way to distinguish lower and higher abilities in L2 oral proficiency. Our qualitative analysis suggests that such differentiation arises from the increasing item length but also in part from specific loci for structural complexity featured in the items.
Reliability of EI Scores

A test of internal consistency (Cronbach’s alpha) was conducted and found to be quite high, $\alpha = .92$, demonstrating impressive reliability by typical measurement standards. This finding is similar to the parallel versions of this EI described in Ortega et al. (1999) and Zhou and Wu (2009), as shown in Table 3.
Table 3. Comparison of Reliability Analysis by EI Test Language

<table>
<thead>
<tr>
<th>EI Test Language</th>
<th>Cronbach’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>English(^a)</td>
<td>0.93</td>
</tr>
<tr>
<td>German(^a)</td>
<td>0.96</td>
</tr>
<tr>
<td>Japanese(^a)</td>
<td>0.95</td>
</tr>
<tr>
<td>Spanish(^a)</td>
<td>0.97</td>
</tr>
<tr>
<td>Chinese(^b)</td>
<td>0.97</td>
</tr>
<tr>
<td>French</td>
<td>0.92</td>
</tr>
</tbody>
</table>

\(^a\)reported by Ortega et al. (2002); \(^b\)reported by Zhou and Wu (2009)

It can be argued that the high reliability comes from three factors. First, a polytomous scoring system is used (from 0 to 4 per item) rather than dichotomous (0-1), allowing for considerable fine-tuning in how performances are scored, and thus creating more variability, which is good for boosting reliability in general. Second, the scoring scale also likely supports consistent scoring by raters because at least 3 of the 5 points on the scale are quite low inference: 0 = no repetition, 1 = fragments, 4 = perfect repetition of both form and meaning. The only two score points that are more difficult to negotiate and at times engender grey areas for coding are 2 and 3 (see Appendix B). Third, by design the test includes 30 items covering a wide range of difficulties suitable for the goals of both challenging and supporting test-takers with a range of proficiencies. Namely, some of the very short first stimuli are likely to be easy for even low proficiencies and some of the longest stimuli are likely to be difficult for even high proficiencies. In addition, as we have shown in the qualitative analysis, varying linguistic challenges are posed independently of syllable length by various grammatical differences across items.

Discussion
The results of the current study suggest that the French EI reported here is a reasonably valid and reliable measure of oral proficiency for use in SLA research. A strong positive correlation was found between the EI scores and (a) $D$ in the oral interview, (b) end-of-year university marks, and (c) speech rate on the oral picture-based narrative. The correlation of EI with $D$ in speaking (rather than writing) performance is logical, as both tap into a similar aspect of proficiency and communication mode that demands integrative speaking and listening skills. Likewise, the university end-of-year marks in the present study were an average of major exams focusing on listening, speaking, and writing, and thus these students were assessed in their language program, to a large extent, on their ability to do listening and speaking tasks in French. A correlation between EI and speech rate on the oral picture-based narrative is expected considering the history of research in language testing demonstrating this relationship (e.g., Iwashita et al., 2008).

Furthermore, our interpretation that this EI is a valid and reliable measure of oral proficiency receives some support from the fact that, in the current study, $D$ does not correlate between writing and speaking (note also that the third highest correlation was found between $D$ in speaking and speech rate in the picture-based narrative, further triangulating this interpretation). This pattern is suggestive of Hulstijn’s (2011, 2012) idea of a BLC and a HLC, with the EI performance, the oral picture-based narrative, and the oral interview performance more clearly drawing from BLC; it may also suggest that different language use is happening in the two modes; or it may be that the mode differences are widened by genre differences between an interview and an essay. In any case, that $D$ for speaking correlates well with EI and with speech rate serves as one source of criterion-related validity evidence for the EI; that these three scores are the highest predictors of marks is likely a good indication that EI taps into something that is close to the kind of proficiency development that we would want to predict in the instructed setting we are investigating in our main study.
The measure of vocabulary recognition, X-lex, does not correlate well with any of the other variables. This finding is most likely due to a truncated range of values; all of the learners scored more or less at the same band. This result could provide a good argument against using this lexical test measure as an indication of individual proficiency development, unless that development is investigated across major time and concomitant learning intervals.

As a reminder, the range of EI scores went from 36-97 (out of a total possible score of 120), thus spanning individual performances between 30% and 81% repetition success. This great variability is noteworthy, given that all 29 learners came from the same institutional level: the end of the second year in this university curriculum. This finding therefore makes a strong case against using mere institutional level as a measure of proficiency, a practice that remains widespread in SLA research (Callies et al., this volume; Thomas, 2006). Where more nuanced distinctions among learners may be needed, a measure like the EI would seem to be particularly useful.

For SLA researchers working in the field of L2 French, in particular, and for those interested in crosslinguistic SLA research more generally, our item analyses uncovered interesting, possibly language-specific sources of difficulty for our French as a foreign language sample (see McManus et al., this volume, for an implementation of this particular EI within a study on the acquisition of the French subjunctive). Beyond the linguistic examples we have detailed in the Results section, we imagine that a comparison of items across different language versions of this particular EI might uncover more language-specific sources of difficulty, including those that may be uniquely French. This would be a fruitful area of future research, particularly for those interested in crosslinguistic SLA. Additionally, this type of analysis could be a way to investigate the effect that language cognates might have on specific item scores. For example, German speakers of English and Portuguese
speakers of Spanish may be at an advantage on certain EI test items if cognates are used (e.g., Van der Slik, 2010).

**Conclusion**

We would like to acknowledge some limitations of the current study. First, because all participants came from the same institutional level, our investigation of the predictive validity of this EI test is limited; in the future we hope to employ the EI with a much wider range of proficiencies spanning all curricular levels. Additional investigations would also look into distinct populations of learners, comparing for example heritage and non-heritage learners (see Wu, 2011, and Zhou, 2011), school and university learners, and so on. Overall, a larger $n$-size and a wide range of proficiencies would be necessary to produce conclusive evidence of the validity of the French EI and of its utility as a tool for measuring L2 French proficiency for the many research purposes demanded in SLA, from recruitment, to assignment to treatments or groups, to contextualization and interpretation of any findings about L2 acquisition (Norris & Ortega, 2012). Assuming a larger sample tapping wider-ranging proficiencies and other salient learner variables, it would be interesting to conduct a cluster analysis on the data and look for natural subgroups which could serve as examples of proficiency level or other factorial differences.

These limitations notwithstanding, the results of the present study highlight the benefits of this parallel EI version. First and foremost, the French EI seems to be a reasonably valid and reliable measure of oral proficiency, demonstrating substantial concurrent validity with lexical diversity in oral interviews, with speech rate on an oral picture-based narrative, and with end-of-year university course marks largely derived from speaking and listening performance. Second, and equally important for research use purposes, it is quick to administer and score. Third, the test itself is inexpensive; however, some kind of recording
device is necessary. Free recording software is available online (e.g., Audacity) making the possession of a separate digital recorder unnecessary. Fourth, based on the present suggestive – if tentative – results, the French EI could be used in conjunction with Tremblay’s (2011) French cloze test to investigate whether differences appear in assessment of learner proficiency based on mode, oral or written, type of knowledge, implicit or explicit (Ellis, 2005), or type of competence along the BLC and HLC distinction (Hulstijn, 2011, 2012). Some research is already underway by Gaillard (in preparation; Gaillard et al., 2011), albeit with a considerably different EI design. Comparing EI tests with differing designs to other types of proficiency tests would prove useful in adding to the debate on defining the construct of proficiency. Last but not least, the existence of parallel versions of our EI test in five other languages (Chinese, English, German, Japanese, and Spanish) allows for crosslinguistic comparisons that have been limited in SLA until now, but which may open new research venues for crosslinguistic SLA programs in the future.

**References**


Gaillard, S., Yi, Y.-S. and Tremblay, A. (2011, September) Implementing an elicited imitation task as a component of a language placement test in French at the university level.
Poster presented at the *Midwest Association of Language Testers and Technology for Second Language Learning Conference*, Aimes, IA.


Elicited imitation as a measure of oral proficiency in L2 French (2014)
Kevin McManus (kmcmanus@psu.edu)


Notes

1 See http://audacity.sourceforge.net/

2 This French EI test will be available for download via IRIS (http://www.iris-database.org)

3 See http://www.lognostics.co.uk/tools/index.htm

4 A Korean version of the same EI test has recently been developed and piloted by Kim and Tracy-Ventura (in preparation).
Appendix A

Elicited Imitation – Newly developed French stimuli, literal English translation, and parallel English stimuli from Ortega et al. (1999), with syllable numbers for the English version in parentheses

1. Je dois aller au coiffeur. (“I have to go to the hairdresser’s”)  
I have to get a haircut (7 syllables)

2. Le livre rouge est sur la table. (“The red book is on the table”)  
The red book is on the table (8 syllables)

3. Les rues sont larges dans cette ville. (“The streets are wide in this city”)  
The streets in this city are wide (8)

4. Il prend une longue douche tous les matins. (“He has a long shower every morning”)  
He takes a shower every morning (9)

5. Qu’as-tu dit que tu vas faire aujourd’hui? (“What did you say that you are going to do today?”)  
What did you say you were doing today? (10)

6. Je ne sais pas s’il sait très bien conduire. (“I don’t know if he can drive all that well”)  
I doubt that he knows how to drive that well (10)

7. Après le repas, j’ai fait une paisible sieste. (“After dinner I had a peaceful nap”)  
After dinner I had a long, peaceful nap (11)

8. Il est possible qu’il se mette à pleuvoir demain. (“It is possible that it might rain tomorrow”)  
It is possible that it will rain tomorrow (12)

9. J’adore les films, surtout ceux qui finissent bien. (“I love films, especially those that have happy endings”)  
I enjoy movies which have a happy ending (12)

10. Les maisons sont très jolies mais inaccessibles. (“The houses are very nice, but inaccessible”)  
The houses are very nice but too expensive (12)

11. Le petit garçon, dont le chaton est mort, est triste. (“The little boy whose kitten died is sad”)  
The little boy whose kitten died yesterday is sad (13)

12. Le restaurant est censé servir de très bons plats. (“The restaurant is supposed to serve very good food”)  
That restaurant is supposed to have very good food (13)

13. Je veux une maison où mes animaux peuvent habiter. (“I want a home where my pets can live”)
Elicited imitation as a measure of oral proficiency in L2 French (2014)
Kevin McManus (kmcmanus@psu.edu)

I want a nice, big house in which my animals can live (14)

14. Tu aimes bien écouter de la musique country, n’est-ce pas? (“You like listening to country music, don’t you?”)
You really enjoy listening to country music, don't you (14)

15. Elle a fini de peindre les murs de son appartement. (“She finished painting the walls in her flat”)
She just finished painting the inside of her apartment (14)

16. Traverse la rue au feu rouge et puis continue juste tout droit. (“Cross the street at the red light and then just continue straight on”)
Cross the street at the light and then just continue straight ahead (15)

17. La personne que je vois a un sens de l’humour fabuleux. (“The person I’m dating has a great sense of humour”)
The person I'm dating has a wonderful sense of humour (15)

18. Elle ne commande que de la viande et ne mange aucun légume. (“She only orders meat and never eats vegetables”)
She only orders meat dishes and never eats vegetables (15/16)

19. J’aimerais que le prix des maisons de ville soit accessible. (“I would like town houses to be more accessible”)
I wish the price of town houses would become affordable (15)

20. J’espère que le printemps arrivera plus tôt que l’an dernier. (“I hope spring arrives sooner than last year”)
I hope it will get warmer sooner this year than it did last year (16)

21. Un de mes meilleurs amis s’occupe des enfants de mon voisin. (“One of my best friends looks after my neighbour’s children”)
A good friend of mine always takes care of my neighbour’s three children (16)

22. Le chat noir que tu as nourri a été chassé par le chien. (“The black cat that you fed was chased by the dog”)
The black cat that you fed yesterday was the one chased by the dog (16)

23. Avant de pouvoir sortir, il doit finir de ranger sa chambre. (“Before going out, he has to finish tidying his room”)
Before he can go outside, he has to finish cleaning his room (16)

24. Je me suis bien amusé lors de notre sortie à l'opéra (“I had a great time when we went to the opera”)
The most fun I've ever had was when we went to the opera (16)

25. Le voleur que la police a arrêté était très grand et mince. (“The thief that the police arrested was very tall and thin”)
The terrible thief whom the police caught was very tall and thin (17)
26. *Pourriez-vous s’il vous plaît me passer le livre qui est sur la table?* (“Would you please pass me the book that is on the table?”)
Would you be so kind as to hand me the book which is on the table? (17)

27. *Le nombre de fumeurs de cigares augmente chaque année.* (“The number of cigar smokers goes up each year”)
The number of people who smoke cigars is increasing every year (17/18)

28. *Je ne sais pas si le train de 11h30 a déjà quitté la gare.* (“I don’t know if the 11.30 train has already left the station”)
I don't know if the 11:30 train has left the station yet (18)

29. *L’examen n’était pas aussi difficile que ce que vous m’aviez dit.* (“The exam was not as difficult as you had said”)
The exam wasn't nearly as difficult as you told me it would be (18)

30. *Il y a énormément d’individus qui ne mangent rien du tout le matin.* (“There are a lot of people who do not eat anything in the mornings”)
There are a lot of people who don’t eat anything at all in the morning (19)
Appendix B

Scoring guidelines for French elicited imitation task (rubric from Ortega et al., 1999)

<table>
<thead>
<tr>
<th>SCORE 0</th>
<th>Criteria</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Nothing (Silence)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Garbled (unintelligible, usually transcribed as XXX)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Minimal repetition, then item abandoned:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Only 1 word repeated</td>
<td>- le voleur bla bla bla la commerce (item 25) (“the thief blah blah blah the shop”)</td>
</tr>
<tr>
<td></td>
<td>- Only 1 content word plus function word(s)</td>
<td>- suivie la rue jusqu’au feu rouge (item 16) (“follow the street until the red light”)</td>
</tr>
<tr>
<td></td>
<td>- Only 1 content word plus function word(s) plus extraneous words that weren’t in the original stimulus</td>
<td>- bla bla dans la ville (item 2) (“blah blah blah in the city”)</td>
</tr>
<tr>
<td></td>
<td>- Only function word(s) repeated</td>
<td>NOTE: with only, just, yet (meaningful adverbs), score 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SCORE 1</th>
<th>Criteria</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• When only about half of idea units are represented in the string but a lot of important information in the original stimulus is left out</td>
<td>- je ne sais pas... conduire (item 6) (“I don’t know… drive”)</td>
</tr>
<tr>
<td></td>
<td>• When barely half of lexical words get repeated and meaningful content results that is unrelated (or opposed) to stimulus, frequently with hesitation markers</td>
<td>- J’adore la film… (item 9) (“I love the movie”)</td>
</tr>
<tr>
<td></td>
<td>• Or when string doesn’t in itself constitute a self-standing sentence with some (targetlike or nontargetlike) meaning (This may happen more often with shorter items, where if only 2 of 3 content words are repeated and no grammatical relation between them is attempted, then score 1)</td>
<td>- quitter la gare (item 28) (“leave the station”)</td>
</tr>
<tr>
<td></td>
<td>• Also when half of a long stimulus is left out, and the sentence produced is incomplete</td>
<td>- XXX dans ce ville (item 3) (“XXX in this city”)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- bla bla bla tu vas faire aujourd’hui (item 5) (“blah blah blah you are going to do today”)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- le petit garçon dans le château est très triste (item 13) (“the little boy in the castle is very sad”)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Je sais passe s’il a un très bien conduire (item 6) (“I know happen if he has a very good drive”)</td>
</tr>
</tbody>
</table>
### SCORE 2

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>• When content of string preserves at least more than half of the idea units in the original stimulus; string in meaningful, and the meaning is close or related to original, but it departs from it in some slight changes in content, which makes content inexact, incomplete, or ambiguous</td>
<td>- <em>J’adore les films, surtout les qui finissent bien</em> (item 9) (“I love movies, especially them that finish well”)</td>
</tr>
<tr>
<td></td>
<td>- <em>J’aimerais que les prix de maison est plus accessible</em> (item 10) (“I would love the house prices is more accessible”)</td>
</tr>
<tr>
<td></td>
<td>- <em>Les voleurs étaient très grand et mince</em> (item 25) (&quot;The thieves were very tall and thin&quot;)</td>
</tr>
<tr>
<td></td>
<td>- <em>Je ne sais pas si le train de 11h00 a déjà quitté la gare.</em> (item 28) (&quot;I don’t know if the 11 o’clock train has already left the station&quot;)</td>
</tr>
<tr>
<td></td>
<td>- <em>La maison est très jolie mais accessible</em> (item 12) (&quot;The house is very pretty but accessible&quot;)</td>
</tr>
<tr>
<td>• Cases of extra marking or more marked morphology should be considered as meaning change. For example, a present tense repeated as past or as future should be scored as meaning change (score 2).</td>
<td>- <em>Je veux une maison ou les animaux peut habiter</em> (item 13) (&quot;I want a house or the pets is able to live&quot;)</td>
</tr>
<tr>
<td>• Similarly, singular/plural differences between stimulus and repeated string change the meaning, not only the grammar (score 2).</td>
<td>- <em>La rue sont larges dans le ville</em> (item 3) (&quot;The street are wide in the city&quot;)</td>
</tr>
<tr>
<td></td>
<td>- <em>Il prend les longues douches tous les matins</em> (item 4) (&quot;He has long showers every morning&quot;)</td>
</tr>
<tr>
<td></td>
<td>- <em>Avant de pouvoir sortir, elle doit finir de ranger sa chambre.</em> (item 16) (&quot;Before going out, she has to finish tidying his room&quot;)</td>
</tr>
<tr>
<td>• Changes of person (il for elle) change the meaning, so score 2</td>
<td></td>
</tr>
</tbody>
</table>

### SCORE 3

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Original, complete meaning is preserved as in the stimulus. Strings which are quite ungrammatical can get a 3 score, as long as exact meaning is preserved. Synonymous substitutions are acceptable.</td>
<td>- <em>Pourriez-vous me passer le livre qui est sur la table?</em> (item 26) (“Would you pass me the book which is on the table?”)</td>
</tr>
<tr>
<td></td>
<td>- <em>Je dois aller à coiffeur</em> (item 1) (&quot;I have to go to hairdresser’s&quot;)</td>
</tr>
<tr>
<td></td>
<td>- <em>J’espère que le printemps arrivera plus tôt que l’année dernière</em> (item 20) (&quot;I hope that spring will arrive sooner than last year&quot;)</td>
</tr>
</tbody>
</table>
• Changes in grammar that don’t affect meaning should be scored as 3. For instance, failure to supply past tense (had>have) and missing articles should be considered grammar change only (score 3).

• Ambiguous changes in grammar that COULD be interpreted as meaning changes from a NS perspective should be scored as 2. That is, as a general principle in case of doubt about whether meaning has changed or not, score 2.

<table>
<thead>
<tr>
<th>SCORE 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Criteria</strong></td>
<td><strong>Examples</strong></td>
</tr>
<tr>
<td>• Exact repetition: String matches stimulus exactly. Both form and meaning are correct without exception or doubt.</td>
<td>- <em>J'espère que le printemps arrive plus tôt que l’an dernier</em> (item 20) (“I hope that spring arrives sooner than last year”)</td>
</tr>
<tr>
<td>• Missing liaison acceptable</td>
<td>- <em>Je dois aller au coiffeur</em> (item 1) (“I have to go to the hairdresser’s”)</td>
</tr>
</tbody>
</table>