BELT+
Brazilian English Language Teaching Journal

ISSN 2178-3640

Porto Alegre, July-December 2015, v. 6, n. 2, p. 142-161

http://dx.doi.org/10.15448/2178-3640.2015.2.21251

ORIGINAL ARTICLE

Elicited Imitation for Brazilian Portuguese

Deryle W. Lonsdale[1], Jarrett Finlinson Lever[1]

[1] Brigham Young University.

ABSTRACT

Elicited imitation (EI) is an approach to measuring oral proficiency that consists of having test takers hear a sentence and repeat the sentence exactly as they heard it. Though indirect in nature, EI has successfully shown to correlate with previously established oral proficiency examinations, such as the Oral Proficiency Interview (OPI) (Lonsdale and Christensen 2014, Matsushita and Lonsdale 2014, Millard 2011, Thompson 2013). This paper discusses the development, administration, and evaluation of an EI test for the Brazilian Portuguese language. We first discuss the relevant background of oral proficiency examination and EI. After presenting the pertinent research questions, we explain the methodology used to develop the EI test, recruit participants, and administer the test. We present the results and analysis and then summarize the findings, limitations, and possible future work.

KEYWORDS: Oral proficiency testing; Brazilian Portuguese; Elicited imitation.

Corresponding Author:
DERYLE W. LONSDALE
<lonz@byu.edu>

1. INTRODUCTION

This paper discusses the development and assessment of an elicited imitation (EI) oral proficiency examination for second language (L2) learners of Brazilian Portuguese (BP). The accuracy of the EI exam is explored by analyzing the correlation between participants' scores from the EI exam and an Oral Proficiency Interview (OPI).

The purpose of the development, administration, and analysis of this test is to assess whether EI can apply to the assessment of BP as it has for other languages such as English (Graham et al. 2008), French (Millard 2011), Spanish (Thompson 2013), Japanese (Matsushita and Lonsdale 2014), and Mandarin Chinese (Wu and Ortega 2013). If a positive correlation exists between the EI test developed in this study and the OPI ratings, this study will not only further validate EI as an assessment tool for evaluating oral proficiency in BP, but will also build greater confidence in EI's capacity to evaluate subjects' implicit knowledge of a second language.

Such a finding in the realm of second language acquisition (SLA) will provide more evidence to the ongoing question about whether EI can effectively evaluate the accuracy of L2 learners' oral proficiency. Since EI was first used for SLA (Naiman 1974), linguists have debated its accuracy. Critics have questioned whether EI actually assesses comprehension, the ability to reconstruct the meaning (Bley-Vroman and Chaudron 1994), or short-term memory (Erlam 2009). Critics have also scrutinized the various factors that influence the production and execution of an EI test (Gillmore and Tharp 1981 & Tomita, Suzuki and Jessop 2009) as well as fidelity to real conversation features (Timothy 1996). Supporters, on the other hand, have found EI to (a) accurately measure implicit knowledge of a language (Erlam 2009); (b) have strong correlations with other oral proficiency examinations and other measures (Henning 1983; Graham et al. 2008; Millard 2011; Matsushita and Lonsdale 2014; Thompson 2013; Moulton 2012; Lonsdale and Millard 2014; Lonsdale and Christensen 2014); and (c) have a great amount of practical benefits not found in other forms (Jessop, Suzuki and Tomita 2007; Radloff 1991). This study thus hopes to provide more evidence to the support of EI as a valid method of evaluating the accuracy of L2 speakers' oral proficiency in BP.

Additionally, positive correlations could lead to a screening process by which any entity that values proficiency in BP and desires an OPI from its candidates, could pre-screen large numbers of subjects with an EI test with more ease and less expense than simply administering the OPI to all candidates. Currently, the OPI costs over $100, takes 20 to 30 minutes to complete, and requires a certified evaluator to administer and evaluate the test. An EI test takes approximately 10 to 15 minutes, entails minimal cost, is accessible via the internet, can be graded by a computer, and correlates closely with the OPI. The future possibilities for EI oral proficiency assessments in BP—especially if combined with other kinds of computerized tests—are great and this study launches that enterprise.

In this paper we utilize the most recent findings in EI theory to develop, administer, and analyze the EI test results. Following traditional EI test development methodology, this included gathering and selecting 84 sentences from a corpus and narrowing that number to 51 items when

considering factors of sentence length, lexical difficulty, and grammar difficulty. We administered the EI test on the campus of Brigham Young University (BYU); Language Testing International© (LTI) administered the OPI. We used correlational methods to analyze the relationship between the EI scores and the OPI ratings; we also used Item Response Theory (IRT) to assess the discriminating power of the EI test.

2. BACKGROUND

Our work builds on several aspects of language testing. In this section we sketch related developments.

2.1. Oral Proficiency Testing

In an increasingly globalized economy and culture, second language acquisition (SLA) has become highly valued for businesses, universities, churches, and governments throughout the world. Consequently considerable emphasis has been placed on developing oral proficiency evaluations that are reliable, accurate, and cost-efficient. Numerous tests have been developed throughout the last century by individual universities, states, and companies for assessing how well second language (L2) learners speak a language. The most widely recognized and accepted form of evaluating oral proficiency in the United States evolved from the work of the Foreign Service Institute's (FSI, an entity of the Interagency Language Roundtable or ILR) work in the 1950's (Council 2007). The FSI's dissatisfaction with their former tests led to the development of a face-to-face conversation exam that was evaluated using a 0 to 5 level scale of verbal proficiency. The FSI test grew in use among various government agencies and its 0 to 5 scale was modified to a scale of 0, 0+, 1+, and so forth until 5. Eventually the American Council on the Teaching of Foreign Languages (ACTFL) collaborated with the Educational Testing Service (ETS) in the 1970's to develop a modified scale that better addressed their needs and created what is now known as the Oral Proficiency Interview or OPI (Peckham, n.d.).

The OPI became available in the early 1980's and has since become the gold standard of L2 assessment (Thompson 2013). Part of the reason the OPI has gained so much acceptance is how the test allows real conversational speech to be evaluated. Nuances such as context and content, pronunciation, sociolinguistic, and pragmatic performance are evaluated in the OPI that other types of tests have difficulty assessing (Cho 2004). Another element that could contribute to the OPI's success is the systematic method the interviewer uses to find and evaluate proficiency ceilings. By altering the difficulty of the conversation, the interviewer is able to ascertain the maximum level of proficiency a person can sustain in conversation. Performing this process during the test enables the interviewer to fine-tune the assessment of proficiency for each individual. Thus, the conversation-based form and the real-time pinpointing of proficiency are elements that have helped the OPI become accepted throughout the world.

Despite the reputation and reliability of the OPI, its need for a highly trained administrator and substantial cost in money and time make it less

than ideal for mass low-stakes assessments for students, armed forces, missionaries, and employees. Efforts to find a reliable, accurate and cost-efficient alternative assessment thus returned to explore the less well-known elicited imitation (EI) theory of the 1960's and 1970's.

2.2. Elicited Imitation Theory

An elicited imitation (EI) test is a simple, indirect method of assessing language ability and proficiency. EI works by having subjects hear a determined item (or sentence) and having them repeat the item back as close as possible to what is heard. The theory of EI proposes that to hear a sentence, process its meaning, and produce an imitation exactly the same in meaning[1] as that given requires the subject (person) to have a level of proficiency in the language being tested equal to that which the item is examining (Bley-Vroman and Chaudron 1994).

Linguists have applied EI to three areas of language study. Its first documented use was for analyzing first language acquisition (Fraser, Bellugi and Brown 1963). It expanded to L2 acquisition (or SLA) (Naiman 1974), and then extended to language disorder assessment in children (Dailey and Boxx 1979). This paper focuses on the use of EI in SLA. Over the course of its 38 years of SLA applications, EI has received both criticism and praise as a proficiency evaluation tool.

Criticism against EI's use in L2 assessment comes principally from EI's indirect nature of addressing oral proficiency. Most detractors question (a) whether EI evaluates comprehension or reconstruction, (b) whether EI evaluates language proficiency or memory, (c) the inherent test structure, and (d) and the lack of real-life language use.

Bley-Vroman and Chaudron (1994) point out how there is uncertainty as to whether EI analyzes the comprehension or the reconstruction ability of a participant. For example, participants with good comprehension will fully understand the item they hear while participants with good reconstruction ability will be able to "reconstruct" the sentence according to their understanding and state it out loud (Lust, Chien and Flynn 1987). Critics claim that research is unsure as to whether EI evaluates comprehension or reconstruction and this could be significant depending on the definition of second language proficiency that one wants to measure. Also, if a participant excelled in comprehension yet had poor reconstruction ability (or vice-versa), the EI would not accurately reflect their language capacities. Thus, the uncertainty of what the EI is measuring causes some critics to distrust EI.

Another principal challenge against EI's validity is whether EI actually measures a person's 'implicit knowledge' of a second language or simply evaluates the strength of his or her short-term (or working) memory (Erlam 2009). Working memory is defined by Erlam as the memory "responsible for both manipulating and temporarily storing information" as opposed to

[1] Slight differences in pronunciation and prosody (or the rhythm/stress/intonation of the utterance) are allowed in the imitation to the degree that they do not entail a meaning different from that of the item. For example, saying [ɛllo] (transcribed phonetically) for the word hello can be understood as a slight pronunciation difference where the [h] is silent; whereas, saying [ʃip] for the word ship is considered an error because the sequence [ʃip] in English represents sheep and not ship.

long-term memory (67). If EI test-takers were simply memorizing everything they heard and parroting the memory back, the EI test would be evaluating only the participants' working memory capacity and functionality rather than their implicit knowledge of a language. Similarly, Bley-Vroman and Chaudron (1994) have noted how L2 learners can accurately imitate sentences with patterns they have not yet mastered. The working memory of EI test participants allowed them to perform better than their implicit knowledge would suggest. This distinction between memory and oral proficiency can be controlled for by how researchers construct the items, yet critics claim that further research is still needed to distinguish what EI is truly testing.

Some linguists have doubted EI due to the numerous factors that can influence EI tests. For example, sentence complexity, sentence rhythm pattern, and information density were suspected of causing discrepancies between Arabic youth and adults speakers learning English (Gillmore and Tharp 1981).There are also several other factors beyond just the test itself that can influence the validity of the EI test, none of which have been researched yet. For example, the design and administration of the test is crucial for all EI tests as are the language recognition and test grading elements of the test (Tomita, Suzuki and Jessop 2009). Thus, the number of possible hidden factors influencing EI causes many to distrust it.

Lastly, critics of EI also see less value in EI because it does not evoke the capacities of real-life language use. McNamara (1996, 31) supports other researchers in stating how any language proficiency test must "replicat[e] reality in the test's 'setting and operation.'" EI, though it may have items taken from corpora of written language or transcribed oral speech, lacks the personable interaction, the spontaneous responses, or specific "language-use situations" that researchers such as McNamara feel are such critical elements of oral proficiency.

Thus, the uncertainty as to whether EI evaluates comprehension or reconstruction, memory or language proficiency, and the lack of real-time language use are three significant factors that disconcert researchers about the "slippery" nature of EI (Vinther 2002, 62) since its use in SLA began (Naiman 1974).

The great need for efficient SLA evaluation has led to recent renewed interest in EI. Research over the years has expanded the knowledge of EI and has found promising results in various aspects that make EI more attractive. Evidence that supports the use of the EI includes (a) the ability of the EI to assess implicit knowledge, (b) the correlation of EI exams with established levels of proficiency, (c) the large number of benefits inherent in EI.

Researchers have conducted EI experiments in different ways to analyze whether EI assesses implicit knowledge of a language. Erlam found (2009) that both native and non-native speakers of English were able to both repeat grammatical sentences and correct ungrammatical sentences in their repetitions. This ability to instinctively correct an ungrammatical sentence one hears supports the belief that EI requires participants to process the test items first and is not evaluating their rote memory capacity. Erlam also found significant correlations between the results of her EI test and the International English Language Testing System (IELTS) test, which provides even further support that her EI test is accurately measuring implicit knowledge.

Other researchers have also found strong correlations between EI tests and established levels of proficiency, which adds credence to the hypothesis that EI can effectively assess oral proficiency. Henning (1983) found that EI tests had the highest degree of validity amongst several factors (such as raw score, fluency, pronunciation, grammar, and a combined fluency-pronunciation-grammar rating) when three different tests (an imitation, an interview, and a completion test) were compared among Egyptian learners of English. Millard (2011) found statistically significant correlations between both human-rated and computer-rated EI and OPI scores for the French language. Research at the Missionary Training Center (MTC) of The Church of Jesus Christ of Latter-day Saints in Provo, Utah found significant correlation between an English EI test developed there and the Language Speaking Assessment (LSA) exam which the MTC has used since 2004 to estimate the fluency of incoming missionaries (Moulton 2012). Many other correlations exist between EI examinations and established oral proficiency tests that further support the use of EI to evaluate oral proficiency.

From a more practical view, EI has several benefits that other oral proficiency evaluation tests do not have. For example, many elements of L2 structure—such as syntax, pronunciation markers, and discourse markers—can be easily elicited with EI (Jessop, Suzuki and Tomita 2007). Jessop et al. also explain how compared to other tests, researchers have more control over the administration and analysis with EI procedures and that the EI test can be used with both children and adults. Some other important benefits include how EI (a) can evaluate a wide range of L2 abilities, (b) can be administered in large-scale testing, and (c) can remain viable even when other people hear the responses of the test taker (Radloff 1991). Thus several benefits to EI make it a viable alternative for evaluating oral proficiency.

In conclusion, evidence has been found that supports EI's (a) ability to evaluate implicit language knowledge, (b) ability to correlate well with already accepted means of examining SLA, and (c) inherent benefits as a testing system. With numerous such experiments performed over the last three decades, more linguists now call EI a "promising assessment technique" that has both validity and reliability (Erlam 2009).

Understanding the long debate between the critics and defenders of EI gives greater understanding of the importance of this study. If this study is successful, it will provide further weight to the validity of EI theory.

2.3. Portuguese EI

As mentioned earlier, EI studies from our research group have found successful results in the recent past with several languages (English, French, Spanish, and Japanese) while as yet no studies apparently address Brazilian Portuguese (BP).

For the English langauge, our researchers analyzed the role of lexical difficulty in EI tests and uncovered interesting results. In one study, Graham et al. (2010) evaluated English EI test results for sentence length (in terms of syllables), lexical frequency (non-corpus based items were created using specific frequency ranges within the British National Corpus (BNC)), lexical density, and morphological density. Their analysis revealed that 73% of the

items' difficulty was attributed to sentence length, 7% was attributed to lexical frequency, and 2% was attributed to lexical density, while morphological complexity was not found to be significant. These findings demonstrate the importance of sentence length in EI items, while additionally lexical difficulty plays as an important role in selecting EI items.

As French and Portuguese are both Romance languages, Millard's work with French supports the belief that an EI test for Brazilian Portuguese would be successful in correlating well with the OPI. Millard found the French EI test he developed correlated well with the OPI when the items were graded by humans ($r = 0.918$) and when graded by computer (0.883). Should the results from the present Portuguese project someday be taken and analyzed by computer, the indications are that a good correlation is probable.

Since Spanish is more similar to Portuguese than French is, recent work in this language is of special interest. Thompson (2013, 46, 51) not only found the Spanish EI test she developed to be a "reliable measure of overall language proficiency" between the computer- and human-graded scores ($r = 0.80$), she also uncovered several interesting and pertinent findings. For example, using Item Response Theory (IRT), she found that the elements of grammar difficulty did not predict test item difficulty.

Another relevant finding of Thompson was that the maximum number of syllables that test takers could generally repeat correctly was 34 syllables (Personal communication, March 6, 2013). It is also relevant that little was known or found about how the lexical difficulty interacted with the difficulty of the test items (Thompson 2013, 48). All of these specific findings in addition to the successful correlation between computer- and hand-graded scores contribute to this study.

Compared to available knowledge about English, French and Spanish EI tests, very little is known concerning the use of either European or Brazilian Portuguese with EI. The literature appears to be absent of documentation about using the Portuguese language in EI tests. This paper thus seeks to fill that gap by developing an EI oral proficiency test using the best resources and methods possible and discover whether the EI test can correlate well with the results of an OPI.

In this study, then, we endeavor to answer the question of how a Brazilian Portuguese (BP) elicited imitation (EI) test could be constructed and administered, and whether it would show a significant correlation with the oral proficiency interview (OPI) ratings for second-language (L2) learners of the language,

3. METHODOLOGY

To produce the working test, we followed methods established by the Pedagogical Software and Speech Technology (PSST) research group of Brigham Young University in Provo. PSST's pattern for developing EI exams is based on ten years of research and is manifest in their work with Spanish, French, Japanese, and English. The foundational elements of said pattern are based on (1) receiving approval from the Institutional Review Board (IRB), (2) developing an EI test, (3) selecting and recruiting participants, (4) administering the test, and (5) analyzing the results.

3.2. EI Test Preparation

We developed 51 EI test items (i.e. sentences) according to three commonly adopted criteria: grammar difficulty, length in syllables, and lexical difficulty (Graham et al. 2010).

Following Millard's (2001) procedure for developing French EI items, we identified key grammar principles for oral BP by utilizing the grammar grid *LS Grammar Grid—Spanish* from the ILR Handbook of Oral Interview Testing for Spanish (Lowe 1982). We first adapted the grammar grid to BP, rearranging certain grammar principles to different difficulty levels and altering a few terms to accurately portray a representative grammar table for L2 learners of BP. The final grammar grid that we constructed and used to select item sentences with from the Portuguese corpus is available online (http://linguistics.byu.edu/thesisdata/ls_grammar_grid_BP.html).

We then extracted 108 sentence items from *O Corpus Do Português* (Davies and Ferreira 2006) that contained grammar elements found in the ILR grammar document. Our corpus queries consisted of vocabulary and grammatical constructions from the various difficulty levels of the grammar grid. For example, to construct an advanced item we searched for sentences exhibiting the most salient and grammatically advanced features at that level (e.g. the ability to use counterfactuals in simple tenses) (Lowe 1982, 3), as seen in the sentence *Se eu tivesse vinte anos até poderia pensar nisso.*

Next we hand-edited the selected corpus items by shortening them to various approximate lengths: short, medium, and long. Of course, we took care to shorten sentences into clauses that—no matter the length—would still remain coherent and stand as complete grammatical sentences. We then removed or replaced all proper nouns with common nouns that would be understood by all L2 learners of BP (e.g., *Rio de Janeiro* was replaced by *cidade)*. The result was a set of 108 items that were based on corpus sentences and were marked for the salient grammatical features they contained. Applying the ILR grid to each sentence, it received a 10-point scale score (0, 1, 1+, 2, 2+, 3, 3+, 4, 4+, 5). While confirming the appropriate grammar difficulty ratings for each item, we eliminated 24 items that were too long, too difficult in grammar and vocabulary, or too archaic in language to be understood by modern BP speakers. We also learned at this time that Thompson had used the same grammar grid we had for her Spanish EI test and had found that the grammaticality of her items did not significantly predict item difficulty (Thompson 2013). Thus, we discarded the prior 10-point scale score we previously gave each sentence and simply counted how many of the salient grammatical features remained in the post-shortened sentences. The grammar-based item creation stage thus yielded a collection of 84 items scored by the number of salient difficult grammatical features they contained (e.g. one item, two items, etc.).

The next step was to record the 84 items with both a male and female native speaker of BP (São Paulo dialect). We recorded the items in a soundproof recording room. Following standard practice, the elicitors were instructed to review the items for a short time before recording them and then to read each item "methodically and clearly" so that every word could be heard (Thompson 2013, 26). The result was 168 high-quality item sound files (84 from each speaker).

From this set of files, a balanced set of 84 items was selected for the test. They spanned the range of items from simple to difficult, and included 42 items from each speaker to avoid a gender effect in item comprehension.

The next step was to assign a syllable count to each item. This is important as Graham et al. (2010) found that sentence length contributes up to 73% of the item difficulty. Syllabification followed the rules listed in the WebCLIPS grammar tutorial (Bateman 2005), taking care to address suspected instances of symalepha (Azevedo 1981, 184-185). Informed by a previous Spanish EI study (Thompson 2013) and due to the proximity of Spanish and Portuguese, we excluded the items with syllable lengths greater than 34, as they would likely be too difficult to repeat. We also eliminated a few items of approximately the same number of syllables at the higher end of the syllable count spectrum.

Finally, we undertook to quantify the items' lexical difficulty. Using the lexical frequencies of the 5,000 most frequent words in BP (Davies and Preto-Bay 2011), we assigned each word in each item a lexical frequency number (LFN) and calculated a mean LFN for each item. Function words were excluded from consideration.

By rating each item based on its length in syllables, grammatical complexity, and lexical frequency, we thus had a gradient of item difficulty. We created and used a system for selecting items from this gradient based on these features, and selected 51 items from the full range of the set of 84 items for inclusion in the test.

3.3. EI Test Administration

To administer the test during the first months of 2014, we recruited a total of 42 volunteer participants from (a) individuals who completed the OPI during the Fall 2013 semester, (b) students in 100- to 600-level Portuguese courses, (c) from native BP speakers, and (d) from non-speakers of BP (i.e. persons who had no training in BP and as little training possible in Spanish and other foreign languages, especially Romance languages). All four categories were recruited from among enrolled BYU students. Classes numbered 100 through 400 roughly correspond to first-year through fourth-year undergraduate classes, and 600-level classes are for graduate students. All three native speakers were born and raised in Brazil. Table 1 summarizes the distribution of all 42 participants by experience level.

Table 1: Distribution of recruited participants

Class level	Number of Participants
Non-speaker	3
100	2
200	4
300	17
400	11
600	2
Native speaker	3
Totals	**42**

We also required that the participants not be Spanish speakers (native or L2) and not be L2 speakers of European Portuguese (EP) speakers. This was to avoid carryover effects from those languages in the testing. A small cash-equivalent incentive was given to the participants on conclusion of the testing.

The independent third party Language Testing International© (LTI) administered the OPI to participants by telephone. Some students took the OPI test before the EI test, and some afterwards.

A proctor adminstered the EI test in a language lab on the BYU campus. Students used high-quality microphone headsets and took a preliminary audio test to assure proper recording. The testing program presented the audio prompt to the participant, had a brief three second pause, and then beeped to notify that the participant may repeat the item. We allotted a time (in seconds) equal to 4 seconds plus the duration of the item for participants to respond to the prompt. Participants could only hear the item once; however, participants were permitted to choose when to advance to the next item by clicking a "Next" button.

Item presentation order involved a gradual increase in item length, then a slight decrease in item length, and then another gradual increase in item length until it surpassed the previous longest item in length. This gradual ascent-decent-ascent pattern was continued through seven tenths of the items. Then, for the remaining three tenths of the test, a similar pattern of decent-ascent-descent item arrangement quickly brought the final item lengths back to the relative lengths of the initial items. The items were sequenced this way in order to (a) present gradually increasingly longer items and (b) avoid priming students to perform better on longer items. We also thought that having the increase-decrease-increase pattern would provide enough variety in length and difficulty that participants would not detect a pattern. Lastly, we placed two of the easiest items at the beginning and two at the end so as to avoid intimidating students at the test onset, and to encourage participant morale on the last items.

3.4. EI Test Grading

Human grading of the EI responses was done via a custom interface developed by the Provo, Utah Missionary Training Center (MTC). A team of five trained students graded the items; four students had extensive training in linguistics whereas the fifth student had a significant amount of Portuguese training. All five graders were native speakers of English and two were L2 speakers of BP while the other three were L2 speakers of Spanish. Since previous EI grading studies have shown that non-native English speakers were able to grade English EI responses without greatly affecting the reliability of the score (Son 2010), we saw no problem in using this team of graders.

Once uploaded to the English Language Center (ELC) server, each item was graded twice. Graders evaluated the participants' responses according to several instructions. First, graders evaluated the participants' audio files according to how many syllables the participant correctly imitated. As seen in Figure 1, the graders selected (and turned green) every syllable that they heard in the participant's audio file.

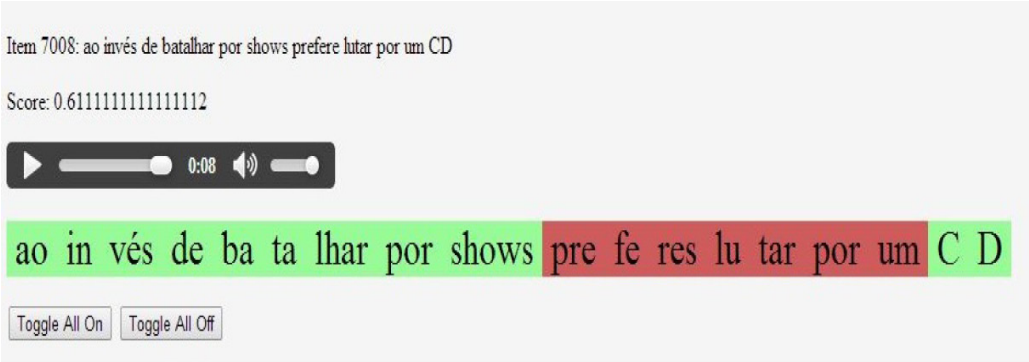


Figure 1: Example of human grading interface

As shown in Figure 1, graders could repeat the audio file as many times as necessary and could see the original prompt's text as well. The syllables that graders selected as correct (green) or incorrect (red) appeared in their orthographic form as we originally divided them (as previously described). We maintained syllables in their orthographic form to (a) economize time in preparing the online grader and (b) allow graders to use their intuition with the orthographic representation as to whether a participant correctly repeated an item rather than use a predefined phonetic transcription that only allows for one interpretation.

To assist the intuition of graders, we provided them with three sources of guidelines for grading. The first two forms are grading guidelines developed by the BYU PSST research group. and are found freely available at the PSST website[2]. These documents contain both general guidelines and specific guidelines for grading EI responses; some of the most salient guidelines are the following:

1. Grade on the syllable level.
2. Never grade pronunciation. As long as you can understand it, count it as correct. If parts of a word are not pronounced (e.g. at the end of a word or elsewhere), do not give full credit: this is a deletion.
3. Transposed words (or syllables) get full credit only for the first word (or syllable).
4. Inserting non-prompt words or repeating words does not count against the participant.
5. Give full credit for contractions and clitics (i.e. the merging of two words).
6. Give full credit for mistakes that are corrected and restarts.
7. If participants repeat a sentence incorrectly (even multiple times), give credit for the last full sentence spoken.
8. Flag overly quiet sound files, unintelligible sound files, or files with extensive background noise accordingly.

Graders were reminded to not grade pronunciation, especially since BP has considerable variation in the pronunciation of word-final vowels and of the phonemes /di/, /de/, /ti/, /te/, word-initial /r/, and word-middle /rr/.

[2] See <http://psst.byu.edu/wiki/index.php/Updated_Grading_FAQS and http://psst.byu.edu/wiki/index.php/Grading_FAQS>.

In regards to participants' OPI scores, Language Testing International© (LTI) assessed each participant's conversation and returned the participant's OPI rating to the BYU's Center for Language Studies (CLS), who then shared those ratings with us via a private spreadsheet.

4. RESULTS

After the participants' scores were double-graded by the five PSST graders and after we had received the OPI ratings from the CLS, we analyzed the data using correlational analysis and Item Response Theory (IRT). Both forms of analyses were performed so as to understand the relationship of the EI scores with the OPI ratings as well as the discriminatory power of the EI test.

4.1. Correlational analyses

We performed four correlational analyses to compare the various human-scored EI and OPI test results. To assure commensurability we first converted participants' OPI ratings into an integer on a 0 to 10 scale[3] (Figure 2a). We assigned the non-speakers a value of 0 (instead of 1) because their responses did not even match the criteria for the Novice-Low rating (American Council on the Teaching of Foreign Languages 2012a)[4]. No participants obtained a Superior or 10 rating so this value does not appear in the remainder of the analysis.

First we ran a two-tailed *t* test to measure correlation between the 42 participants' numerical OPI ratings and human-graded syllable percent score from round 1 (H-EI-1). The result ($r = 0.93$, $p = 0.0000$, $R^2 = 0.8729$) showed a strong correlation between the two tests. The expected range of sample *t*'s was -2.02 to 2.02 and the actual *t*-value was 16.57. The confidence interval of the true population correlation, ϱ, ranged from 0.88 to 0.96. Furthermore, plotting the correlation showed a clear division (a 20% gap on EI syllable scores) between the non-speakers and the lowest performing L2 BP speakers in their EI test scores.

OPI Rating		Numerical OPI Rating
Superior	S	10
Advanced-High	AH	9
Advanced-Mid	AM	8
Advanced-Low	AL	7
Intermediate-High	IH	6
Intermediate-Mid	IM	5
Intermediate-Low	IL	4
Novice-High	NH	3
Novice-Mid	NM	2
Novice-Low	NL	1
Non-speaker	NS	0

Correlational Analysis	*r*-value	R^2 value
H-EI-1 vs. OPI	0.93	0.8729
H-EI-2 vs. OPI	0.92	0.8534
H-EI-(mean) vs. OPI	0.93	0.8707
H-EI-1 vs. H-EI-2	0.98	0.9638

Figure 2: (**a**) OPI ratings conversion (left) and (**b**) OPI vs. EI test score correlations (right).

[3] This linear 0 to 10 scale for categorizing OPI test numerical values is often used by researchers but inconsistent with the ACTFL's inverted triangle definition of the different difficulty levels of each OPI rating. Their model stipulates that an increasing amount of fluency is required for test takers to progress from each rating to the next (American Council on the Teaching of Foreign Languages 2012b).

[4] By definition of the ACTFL, the non-speaking participants in this study would be assumed to have a Novice-Low (or numerical score of 1) proficiency on the OPI. This definition was not utilized in this study.

After the completion of the second round of human grading of the EI test responses, we performed a correlation between participants' OPI ratings and their second-round EI scores. Again there was a strong correlation ($r = 0.92$, $p = 0.0000$, $R^2 = 0.8534$). The expected range of sample *t*'s was -2.02 to 2.02 and the actual *t*-value was 15.26. The confidence interval of the true population correlation, ϱ, ranged from 0.86 to 0.96.

To verify the reliability of the human scores, we performed a correlation between the first and second rounds of human-graded scores. This correlational analysis answers the question "How likely are two different human graders to give the same participant the same score?" A high value would in fact show that the human-graded score is a reliable measure of each participant's proficiency. The correlation was strong ($r = 0.98$, $p = 0.0000$, $R^2 = 0.9638$). The expected range of sample *t*'s was -2.02 to 2.02 and the actual *t*-value was 32.64. The confidence interval of the true population correlation, ϱ, ranged from 0.97 to 0.99.

Finally we calculated the correlation between the OPI scores and the mean of the two human scoring rounds. Again there was a strong correlation ($r = 0.93$, p=0.0000, $R^2 = 0.8707$) with the expected range of sample *t*'s from -2.02 to 2.02 and the actual *t*-value at 16.41.

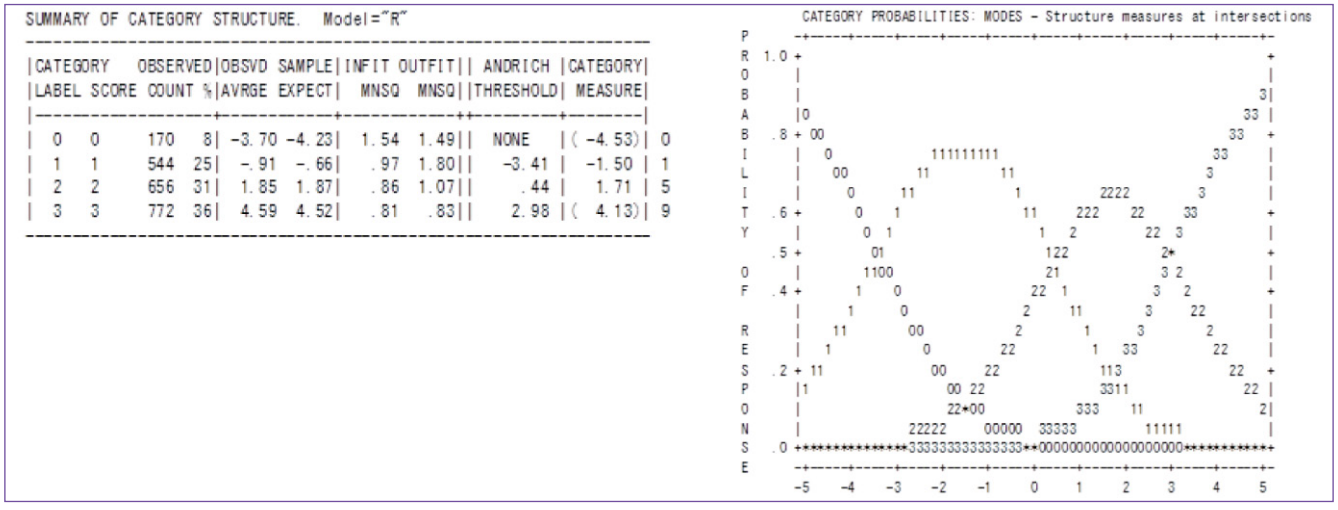
Figure 2b summarizes the various correlation values just discussed.

4.2. Item Response Theory Analysis

We also performed an Item Response Theory (IRT) Analysis on the 51 test items using the one parameter Rasch-model (Graham et al. 2010). We used the WinSteps program for analysis and output generation (Linacre 2014b). The IRT analysis provides:

a) results that are independent of the group that took this test (allowing direct comparison with later groups that might take this same test);
b) results that are independent of these specific Portuguese items (which allows the items to be calibrated and compared with other items).

The next few figures and associated narrative present these results. A more thorough account is available elsewhere (Lever 2014).



SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	ANDRICH THRESHOLD	CATEGORY MEASURE	
0	0	170	8	-3.70	-4.23	1.54	1.49	NONE	(-4.53)	0
1	1	544	25	-.91	-.66	.97	1.80	-3.41	-1.50	1
2	2	656	31	1.85	1.87	.86	1.07	.44	1.71	5
3	3	772	36	4.59	4.52	.81	.83	2.98	(4.13)	9

Figure 3: Summary of Category Structure

Figure 3 plots the structure of the category grading for responses. The numbers in the graph represent a rating of the items (0=no syllables correct, 1=less than half of syllables correct, 2=more than half of syllables correct, 3=all syllables correct). It shows that participants with the least amount of ability (e.g. a person with the lowest logit score—at the left of the x-axis—of less than -3.41) will with highest probabability (over 80% on the y-axis) receive 0 rating (no syllables correct) for an item. On the other hand, participants with an upper-mid range of ability (0.44 to 2.98 on the x-axis) are more likely to get a rating of 2. The categorical separation is clear and well-behaved.

The IRT summary table in Figure 4 shows (in the bottom right corner) that the EI test as a whole has a Person Relability value of 0.98 and an Item Reliability score of 0.98. The Person Reliability value is an indication of how well the test items can measure and distinguish the test-takers from each other. For example, if the same person were tested on these items twice (without remembering the first test experience when tested the second time), a high Person Reliability value indicates a very high probability that he or she would perform similarly on both test experiences. In other words, with a large Person Reliability value, the test is very consistent internally. The large Item reliability indicates that there was a wide range of item difficulty and that there was a relatively large population for the test.

The Person table also shows a separation value of 7.54. This means that the test items naturally divided participants into 7.54 groups of distinct levels of performance.

```
     SUMMARY OF 42 MEASURED Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      99.3      51.0        1.70     .29       .96    -.2   1.25     .1 |
| S.D.      29.8        .0        2.31     .04       .26    1.2   1.05    1.6 |
| MAX.     145.0      51.0        6.08     .44      1.79    3.0   5.02    5.8 |
| MIN.      19.0      51.0       -4.63     .26       .58   -2.5    .38   -2.6 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .30 TRUE SD    2.29  SEPARATION  7.54  Person RELIABILITY  .98 |
|MODEL RMSE    .29 TRUE SD    2.29  SEPARATION  7.88  Person RELIABILITY  .98 |
| S.E. OF Person MEAN = .36                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = 1.00
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .98

     SUMMARY OF 51 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      81.8      42.0         .00     .32       .97    -.2   1.25     .2 |
| S.D.      23.8        .0        2.15     .06       .21     .9   1.06    1.2 |
| MAX.     117.0      42.0        4.16     .51      1.63    1.9   6.05    4.9 |
| MIN.      35.0      42.0       -3.99     .28       .49   -2.8    .32   -2.9 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .34 TRUE SD    2.13  SEPARATION  6.27  Item   RELIABILITY  .98 |
|MODEL RMSE    .33 TRUE SD    2.13  SEPARATION  6.53  Item   RELIABILITY  .98 |
| S.E. OF Item MEAN = .30                                                     |
-------------------------------------------------------------------------------
Item RAW SCORE-TO-MEASURE CORRELATION = -.99
2142 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 2856.68 with 2048 d.f. p=.0000
Global Root-Mean-Square Residual (excluding extreme scores): .4847
UMEAN=.0000 USCALE=1.0000
```

Figure 4: IRT Summary Table

Figure 5a shows the EI scores compared to the OPI ratings. The participants were encoded with OPI ratings (see right-most column) and a Person Subtotal analysis was conducted. The second column from the left shows how the Mean Measure of a group's performance steadily increases with the OPI rating. This fact shows that participants' performance on the EI test correlates well with the OPI rating. The fourth column from the left (Observed Standard Deviation) shows how the standard deviation of group 06 (1.17) is large enough to blur the distinction between group 06's (Intermediate-High rating) performance (Mean Measure 1.73) and group 07's (Advanced-Low rating) performance (Mean Measure 2.12).

The groups were also compared using Welch's T-test, chosen because it does not require equal variances between the groups. Figure 5b shows that the *p*-value (0.507, right-most column) between OPI levels 06 and 07 (Intermediate-High and Advanced Low) is large enough to indicate no significant difference between EI scores of people who achieved the IH and AL ratings on the OPI. Hence these two grous could be combined into one with minimal perturbation. The 06 and 07 groups have the least distinction (p=0.507), followed by the 06 and 08 groups having the next least distinction (p=0.076), with the 05 and 06 groups having the smallest distinction (p=0.016).

Though we do not provide the details here, an Item Statistics:Misfit Order analysis showed which items performed in the least expected manner. Items 20-2, 30-3, 11-3, 12-3, 24-3, 26-3, 18-1, and 24-2 were identified as problematic and hence need to be further examined and possibly removed from the item pool.

Person COUNT	MEAN MEASURE	S.E. MEAN	OBSERVED S.D.	MEDIAN	MODEL SEPARATION	MODEL RELIABILITY	CODE
42	1.70	.36	2.31	2.11	7.88	.98	**
3	-4.40	.23	.32	-4.63	.00	.00	00
1	-1.23	-	.00	-1.23	.00	.00	04
5	-.22	.17	.33	-.21	.83	.41	05
6	1.73	.52	1.17	1.75	4.21	.95	06
15	2.12	.14	.53	2.14	1.69	.74	07
8	3.04	.38	1.01	2.75	3.29	.92	08
4	5.08	.38	.65	4.98	1.42	.67	09

Person CODE	Person CODE	MEAN DIFFERENCE MEASURE	S.E.	t	Welch-2sided d.f.	Prob.
00	05	-4.18	.28	-14.72	4	.000
00	06	-6.14	.57	-10.74	6	.000
00	07	-6.52	.27	-24.16	3	.000
00	08	-7.44	.45	-16.65	8	.000
00	09	-9.48	.44	-21.47	4	.000
05	06	-1.96	.55	-3.57	5	.016
05	07	-2.35	.22	-10.70	10	.000
05	08	-3.26	.42	-7.80	9	.000
05	09	-5.30	.41	-12.85	4	.000
06	07	-.39	.54	-.71	5	.507
06	08	-1.30	.65	-2.01	9	.076
06	09	-3.35	.65	-5.19	7	.001
07	08	-.91	.41	-2.24	8	.056
07	09	-2.96	.40	-7.33	3	.005
08	09	-2.04	.54	-3.80	8	.005

Figure 5: (a) Personal subtotal (left) and (b) Group subtotal (right) analyses

Finally, the Person-Item Map in Figure 6 plots performance of the individual participants (left-hand side) with the items' evaluated performance (right-hand side). The first two numbers of the participants' identification represents the numerical value of the OPI rating they received (see Figure 2a for the numerical to categorical conversion). The second pair of numbers identifies the person. The items' identification represents the difficulty level we originally assigned to the item. The digits preceding the dash represent the numbers of syllables in the item; the digits following the dash represent

the relative Lexical Frequency Number (LFN) difficulty per each sentence length group (3=highest, 2=median, 1=lowest). Persons and items are on the same scale, thus a person with a logit measure of 0 (e.g. participant 05-25) would have a 50-50 chance of correctly answering an item with a logit of 0 (e.g. item 17-1). Since this is a four-point rating scale, the item difficulties on this map indicate the average difficulty response categories rather than any specific category.

The left-(Person) side of the map plots how participants were discriminated from each other by the 51-item EI test, with Level 09 (Advanced-High) participants at the top and the level 00 (Non-Speakers) at the bottom as expected. Overall the distribution is spread in a fairly consistent manner with the associated OPI ratings. Only a few outliers (e.g. 06-35, 06-33, and 06-30) exist. The EI test discriminates well among participant groups.

The right (Item) side of the map illustrates a general trend that the shorter items are generally near the bottom and the longer items generally occur near the top. This pattern agrees with prior findings that item length is the largest contributor to item difficulty. A pattern for the lexical difficulty does exist but is not nearly as strong as the item length pattern.

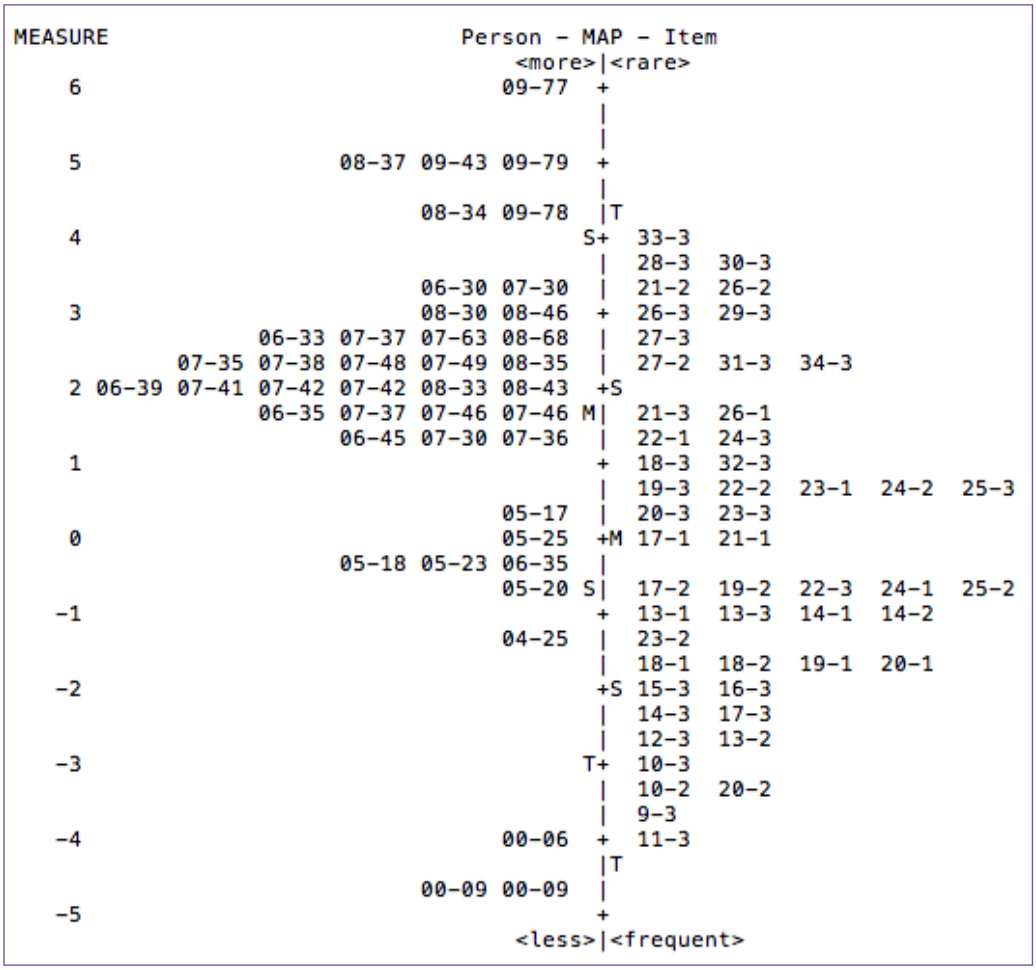


Figure 6: Person-Item map

5. DISCUSSION

In this study we developed and administered an oral proficiency test for Brazilian Portuguese (BP) based on elicited imitation (EI) theory, and then compared its results with scores from an oral proficiency interview (OPI). Our 51-item EI test, when scored by humans, showed a significant and strong correlation with the oral proficiency interview OPI ratings.

These findings also give further credence to using lexical frequency as a means of determining lexical difficulty in EI items. Further statistical analyses will have to be performed to determine the degree to which the lexical difficulty process employed in this study contributed to the overall item difficulty.

The limited number of 42 participants is a sizeable number of participants for an early EI study in a new language, yet still a relatively small considering the population of BP speakers and learners. to A more even representation across the spectrum of BP L2 oral proficiency would also have been desirable.

Employing ACTFL's definition of non-speakers and their definition of the difficulty between OPI ratings would increase future correlational analyses with any EI test that is compared to the OPI. For example, in post-hoc analysis, we found that identifying non-speakers as a Novice-Low or "1" OPI rating increased the correlational *r*-value between the OPI Ratings and H-EI Syllable % Score (Round 1) by 0.01 and slightly increased the R^2 and *t*-values of the three OPI Ratings vs. H-EI Syllable % Score analyses. Future work should also follow Meredith's (1990) findings concerning OPI ratings and various numerical scales. By experimenting with various non-linear numerical scales and OPI categorical ratings, Meredith found a 12.1 increase in R^2 when he evaluated subjects with a non-linear scale where the strength of subjects' performance (i.e. strong, average, or weak) was noted and where large distances were given between each sub-category at the higher levels. Thus, incorporating ACTFL's definition of non-speakers into the numerical scale and Meredith's (1990) findings future work should find higher correlation values for EI tests.

As using lexical frequency as a measurement of lexical difficulty was an exploratory step in this study, future work could research different aspects of lexical frequency (e.g. range, average, median, or others) to determine which measurement gives EI items the greatest discriminating power. Researchers could address this test's limitations, and then form various tests by selecting from the 84 recorded items with different lexical strategies. Doing so would hopefully show which aspect of lexical frequency (average, range, median, etc.) provides the most effective quantification of lexical difficulty.

Besides developing EI tests for other languages, we have also automated the scoring of these tests. Through the use of speech recognition technology and forced alignment, recorded responses can be compared syllable-by-syllable and rated according to pre-set criteria. Since Portuguese language models are available for speech recognition—a prerequisite for such implementations—we expect that development of a computerized EI test for BP should be straightforward.

Lastly, future work could follow the examples of Matsushita and Christensen to develop a simulated speech (SS) test to accompany the EI exam. As the EI exam measures accuracy and the SS test measures fluency, applying both in a study and combining their scores would provide a more holistic assessment for oral proficiency that includes both accuracy and fluency.

ACKNOWLEDGEMENTS

We are grateful to Dr. Blair Bateman of Brigham Young University for advisement on pedagogical aspects of BP oral proficiency. We also appreciate the support of the BYU Center for Language Studies for funding the OPI tests.

REFERENCES

American Council on the Teaching of Foreign Languages. 2012a. *ACTFL Proficiency Guidelines 2012*. American Council on the Teaching of Foreign Languages. Accessed January 29, 2014. <http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf>.

_____. 2012b. General Preface to the *ACTFL Proficiency Guidelines 2012*. Accessed April 28, 2014. <http://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012>.

Azevedo, Milton Mariano. 1981. *A Contrastive Phonology of Portuguese and English*. Washington, D.C.: Georgetown University Press.

Barreiro, Anabela, Luzia Helena Wittmann, and Maria de Jesus Pereira. 1996. Lexical differences between European and Brazilian Portuguese. *INESC Journal of Research and Development* 5 (2), p. 75-101.

Bateman, Blair. 2005. "Divisão Silábica." *BYU WebCLIPS*. Brigham Young University. Accessed May 15, 2013. <http://webclips.byu.edu/>.

Bley-Vroman, Robert, and Craig Chaudron. 1994. "Elicited imitation as a measure of second language competence." In *Research methodology in second language acquisition* 2, p. 245-261.

Cho, Sungdai. 2004. "Oral Proficiency Interview: Pros and Cons." American Association of Teachers of Korean. Durham. p. 144-150.

Council, National Research. 2007. "Appendix D: A Brief History of Foreign Language Assesment in the United States." In *International Education and Foreign Languages: Keys to Securing America's Future*, by Committee to Review the Title VI and Fulbright-Hays International Education Programs, edited by Janet L. Norwood and Mary Ellen O'Connell (p. 360-364). Washington, D.C.: The National Academies Press.

Dailey, K., and J. Boxx. 1979. "A comparison of three imitative tests of expressive language and a spontaneous language sample." Language, Speech, and Hearing Services in the Schools, p. 6-13.

Davies, Mark, and Ana Maria Raposo Preto-Bay. 2011. *A Frequency Dictionary of Portuguese*. CD. Routledge: Taylor & Francis Group.

Davies, Mark, and Michael J. Ferreira. 2006. *Corpus do Português*. Accessed January 1, 2012. <http://www.corpusdoportugues.org/>.

Erlam, Rosemary. 2009. "The Elicited Oral Imitation Test as a Measure of Implicit Knowledge." In Rod Ellis, Shawn Loewen, Catherine Elder, Rosemary Erlam, Jenefer Philp and Hayo Reinders. *Implicit and Explicit Knowledge in Second Language Learning, Testing and Teaching* (p. 65-99). New York: Multilingual Matters.

Fraser, C., U. Bellugi, and R. Brown. 1963. "Control of grammar in imitation, comprehension, and production." *Journal of Verbal Learning and Verbal Behavior* 2, p. 121-135.

Gillmore, R, and R. G. Tharp. 1981. "The interpretation of elicited sentence imitation in a standardized context." *Language Learning* 31 (2), p. 369-392.

Graham, C. Ray, Deryle Lonsdale, Casey Redd Kennigton, Aaron Johnson, and Jeremiah McGhee. 2008. "Elicited Imitation as an Oral Proficiency Measure with ASR Scoring." *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC). Marrakech, p. 1604-1610.

Graham, C. Ray, Jeremiah McGhee, and Ben Millard. 2010. "The Role of Lexical Choice in Elicited Imitation Item Difficulty." Edited by Matthew T. Prior et al. *Selected Proceedings of the 2008 Second Language Research Forum:* Exploring SLA perspectives, positions, and practices, p. 57-72.

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

Jessop, Lorena, Wataru Suzuki, and Yasuyo Tomita. 2007. "Elicited Imitation in Second Language Acquisition Research ." *The Canadian Modern Language Review* / La Revue canadienne des langues vivantes 64 (1), p. 215-238.

Lever, Jarrett Finlinson. 2014. "Elicited Imitation Test for Brazilian Portuguese", BA Honors Thesis; Provo, UT: Brigham Young University.

Linacre, J. M. 2014a. Winsteps® Rasch measurement computer program. Accessed April 15, 2014. <Winsteps.com>.

—. 2014b. Winsteps® Rasch measurement computer program User's Guide. Accessed April 15, 2014. <Winsteps.com>.

Lonsdale, Deryle, and Benjamin Millard. 2014. "Student achievement and French sentence repetition test scores." *Proceedings of the Ninth Language Resources and Evaluation Conference* (LREC) 2014. Reykjavik, p. 2719-2725.

Lonsdale, Deryle, and Carl Christensen. 2014. "Combining elicited imitation and fluency features for oral proficiency measurement." *Proceedings of the Ninth Language Resources and Evaluation Conference* (LREC). Reykjavik. p. 1956-1961.

Lowe, P. 1982. *ILR Handbook on Oral Interview Testing*.

Lust, B., Y. Chien, and S. Flynn. 1987. What children know: methods for the study of first language acquisition. In B. Lust. *Studies in the acquisition of anaphora* (Vol. II, p. 271-356). Dordrecht: D. Reidel Publishing Company.

Matsushita, Hitokazu, and Deryle Lonsdale. 2014. "How to use simulated speech to assess learner Japanese oral proficiency." In Jeffrey Connor-Linton and Luke Wander Amoroso (eds.), *Measured Language: Quantitative Studies of Acquisition, Assessment, and Variation* (p. 171-182). Washington, D.C.: Georgetown University Press.

Meredith, R. Alan. 1990. "The Oral Proficiency Interview in Real Life: Sharpening the Scale." *The Modern Language Journal* 74 (3), p. 288-296.

Millard, Benjamin. 2011. *Oral Proficiency Assessment of French Using an Elicited Imitation Test and Automatic Speech Recognition*. MA Thesis, Provo, UT: Brigham Young University.

Moulton, Sarah. 2012. *Elicited imitation testing as a measure of oral language proficiency at the Missionary Training Center*. MA Thesis, Provo, UT: Brigham Young University.

Naiman, N. 1974. "The use of elicited imitation in second language acquisition research." *Working Papers in Bilingualism* 2, p. 1-37.

Peckham, Bob. n.d. The Interagency Language Roundtable Scale: Introduction from CALL. Accessed April 9, 2014. <http://www.utm.edu/staff/globeg/ilrhome.shtml>.

Radloff, C. 1991. *Sentence Repetition Testing: For Studies of Community Bilingualism*. Dallas: Summer Institute of Linguistics and the University of Texas at Arlington Publications in Linguistics.

Son, Minhye. 2010. *Examining Rater Bias: An Evaluation of Possible Factors Influencing Elicited Imitation Ratings*. MA Thesis, Provo, UT: Brigham Young University.

Thompson, Carrie A. 2013. *The Development and Validation of a Spanish Elicited Imitation Test of Oral Language Proficiency for the Missionary Training Center.* PhD Dissertation, Provo, UT: Brigham Young University.

Timothy, Francis McNamara. 1996. *Measuring Second Language Performance.* New York: Longman.

Tomita, Yasuyo, Wataru Suzuki, and Lorena Jessop. 2009. "Elicited Imitation: Toward Valid Procedures to Measure Implicit Second Language Grammatical Knowledge." *TESOL Quarterly* 43, p. 345-350.

Vinther, Thora. 2002. "Elicited imitation: a brief overview." *International Journal of Applied Linguistics* 12 (1), p. 54-73.

Weber, Jerome C., and David R. Lamb. 1970. *Statistics and Research in Physical Education.* Saint Louis: The C. V. Mosby Company.

Wu, Shu-Ling, and Lourdes Ortega. 2013. "Measuring Global Oral Proficiency in SLA Research: A New Elicited Imitation Test of L2 Chinese." *Foreign Language Annals* 46, p. 680-704.

Submitted: 04/07/2015
Accepted: 03/11/2015