2020-04-13

Development and Validation of a Portuguese Elicited Imitation Test

Braden Beldon Reynolds
Brigham Young University

The Development and Validation of a Portuguese Elicited Imitation Test

Braden B. Reynolds

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

Master of Arts

Blair E. Bateman, Chair
Michael Child
Nieves P. Knapp

Department of Spanish and Portuguese

Brigham Young University

ABSTRACT

The Development and Validation of a Portuguese Elicited Imitation Test

Braden B. Reynolds
Department of Spanish and Portuguese, BYU
Master of Arts

Elicited imitation (EI) is a method of assessing oral proficiency in which the examinee listens to a prompt and attempts to repeat it back exactly as it was heard. Research over recent decades has successfully established correlation between EI testing and other oral proficiency tests, such as the Oral Proficiency Interview (OPI) and the OPI by computer (OPIc). This paper details the history of oral proficiency assessment as well as that of EI. It then outlines the development process and validation of a Portuguese Elicited Imitation test. The processes of item selection and item validation are detailed followed by the criterion-related validation through a statistical correlation analysis of participants' results on an official American Council on the Teaching of Foreign Languages (ACTFL) OPIc and their predicted OPIc scores which were based on their results of the Portuguese EI calibration test. Results of the statistical analysis revealed a strong correlation between the predicted scores of the EI test and the actual OPIc scores. In order to go beyond previously completed EI research, this paper addresses the issue of face validity which has been a challenge for the proliferation of EI testing. Analysis of a survey administered after participants' completion of the two tests (OPIc and EI) addresses the experiences and reactions of the participants to the two testing formats. Suggestions for future use of EI as well as future research will be presented.

Keywords: elicited imitation, oral proficiency assessment, Oral Proficiency Interview-computerized, face validity.

ACKNOWLEDGEMENTS

I would like to start by thanking my committee chair, Dr. Blair Bateman, as well as the other members of my committee, Dr. Michael Child and Dr. Nieves Knapp. Their accessibility, recommendations, and direction have been invaluable to the success of this project. I also want to thank my friends at Emmersion Learning for trusting me with this project, particularly Jacob Burdis and Dani Houston. I would also like to thank Zac Mayne for his statistical support and Judson Hart for his advice and collaboration during the final stages of my thesis.

I would also like to thank Lohany Cruz Bakr, Rafael Lopes, and Ruth Baptista for loaning me their voices in the recording process. I want to thank Marissa Jensen for her assistance in the rating process and all of the participants who took the time to complete the tests as well as the survey.

Finally, I would like to thank my family and particularly my wife, Julie, who has supported me throughout this program. Without her encouragement and love I would not have been able to complete this project.

TABLE OF CONTENTS

CHAPTER 1

Introduction

The motivations for identifying a second language learner's ability to actually communicate proficiently in the target language are varied. Companies around the world need employees who are capable of serving their interests in foreign markets. Militaries and defense organizations are in constant need of multi-lingual operators for any number of reasons ranging from intelligence gathering to field interpretation to international relationship building. More applicable to the populace of this university is the need for young members of The Church of Jesus Christ of Latter-Day Saints to learn a foreign language in order to serve throughout the world in a religious missionary capacity. In many cases, the learner's oral proficiency level needs to be understood for the simple purpose of continuing the language teaching and learning process most efficiently and effectively. The other skills of writing, reading, and listening are comparatively straight forward in their assessment processes since they can be assessed through written tests that are graded as either correct or incorrect. However, assessing oral proficiency continues to be theoretically problematic for educators and researchers. Historically, and even in today's classroom, success in most courses is often linked to achievement in reading and listening comprehension as well as writing ability. In many cases oral proficiency is cited as a goal of a given course or program, but less commonly addressed as one of the measures of achievement within the curriculum. This is partially because definitions of oral proficiency are as loose and diverse as oral proficiency researchers themselves.

However, the need for accurate tools of assessment is only increasing as the world becomes more connected and demand for and accessibility to foreign language education increases. Bowden (2016) reviewed multiple surveys and studies conducted on how second

language proficiency is measured and noted the myriad methods for estimating proficiency and in many cases, a lack of measurement. Similarly, Leclerq and Edmonds (2014) commented that in their review of L2 proficiency research articles, multiple researchers lamented the lack of careful definition or empirical evidence to support proficiency level assignments applied to learners.

One of the primary characteristics and challenges of oral proficiency testing is the involvement of interviewers or raters. As previously mentioned, the general assumption is that testing oral proficiency inherently includes some version of an oral interaction between an evaluator or rater and a learner. Because raters come from a variety of backgrounds, their assessment methods may be varied. Chalhoub-Deville (1995) cited a number of studies in which different groups of raters were compared. These groups consisted of non-native and native speakers of the target language, both living in and outside of communities where the target language is spoken. Comparisons were also made with those who were teachers and non-teachers and who fell within the previously mentioned categories. There were some consistencies such as the tendency of non-native speakers to focus on grammar, with natives focusing on fluency and pronunciation. However, no general conclusions could be drawn from the different studies cited, with the exception of the observation that there are clear inconsistencies and further research required to further define the impact of varying rater backgrounds on oral proficiency assessment outcomes. Rater involvement in proficiency assessment, particularly in research, comes with several challenges that hinder access to accurate oral proficiency assessment. Some of these challenges include cost of training, cost of administering, and logistics (Drackert, 2015). This project will attempt to address these burdens associated with oral proficiency assessment directly through the development of an elicited imitation test which can serve as a much cheaper method.

Oral Proficiency Rating Scales

For decades organizations, including governments and international bodies, have attempted to come up with a reliable way to assess oral proficiency. Prior to discussing the two principal oral proficiency assessment systems used in the U.S., it is worthwhile mentioning the Common European Framework of Reference (CEFR), which is the internationally recognized standard in Europe for language assessment, comparable to the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines. Over the last decade the two organizations have made efforts to align the two rating systems in order to allow for CEFR ratings on ACTFL proficiency assessments (American Council on the Teaching of Foreign Languages, n.d.). This project, however, will maintain a U.S. focus.

In the 1950s, the U.S. government identified the need to document and categorize the foreign language capability that existed within the armed forces. After several years of development, they created a six-point scale that was used to assign a specific proficiency level to the speaker. Over time the scale was revised, and the testing instrument was customized to fit the needs of different federal organizations. The scale was modified to include "plus" levels inside each of the broader levels (1-6). By 1985 the now 11-point scale (0,0+,1,1+,2,2+…5) had been edited and fine-tuned under the direction of the Interagency Language Roundtable (ILR) and was thereafter labeled the "ILR Scale." This scale is now used to assess any student coming out of a federal language course as well as many others.

The Oral Proficiency Interview

In order to identify where a language learner is at on any of these scales, a process called the Oral Proficiency Interview (OPI) was developed. As alluded to in the introduction, the OPI consists of a one-on-one conversation between a rater and an examinee. It is an adaptive

interview, meaning that the course of questions adjusts to the interests and capabilities of the interviewee. Generally, these interviews last no longer than half an hour. Because the government has to administer about 3500 of these OPIs with trained raters every year, it is necessary to have several hundred raters trained to administer the interview (Herzog, 2003), which constitutes a significant financial and logistical burden.

In addition to the government, other parties have maintained an interest in the creation of a reliable system of oral proficiency assessment. Academic institutions and foreign language educators have payed close attention to the developments within government programs regarding the ILR. The government's increasing need for reliable language testing eventually required the assistance of outside language professionals, many of whom were university professors (Liskin-Gasparro, 1984). The ILR became a tool familiar to language educators and the result was a modified scale that included the above mentioned "+" sublevels as they recognized the need to more clearly define the abilities of their students. The focus on oral proficiency continued to increase throughout the 1970s by both government officials as well as educators, with the latter seeking an instrument comparable to that of the government ILR by which to assess oral proficiency, coining the term "common yardstick" in reference to the future product. Eventually a grant from the government to ACTFL resulted in a more descriptive set of guidelines first published in 1982 (Omaggio, 1986). These new guidelines replaced the ILR numerical scale with the levels of *novice, intermediate, advanced,* and *superior,* with the first three levels eventually being subdivided into sublevels *low, mid,* and *high.* Each of these sublevels was accompanied by a robust description of the communicative competencies of an individual at that level.

For the past thirty years the government has worked directly with ACTFL to ensure that their OPI scales complement one another (Herzog, 2003). Like every other oral proficiency

model, the ACTFL Proficiency Guidelines have been intensely scrutinized and critiqued. Particularly in the 1980s and 90s, multiple researchers published investigations of validity (e.g. Dandonoli & Henning, 1990) and critiques of definitions and methods (e.g. Bachman & Savignon, 1986; Lantolf & Frawley, 1985). However, 35 years after their inception, the term OPI, when used amongst language researchers, is nearly always used in reference to the ACTFL Proficiency Guidelines (Liskin-Gasparro, 2003). The OPI has been administered thousands of times over the last several decades, in spite of some of the challenges associated with its use.

Oral Proficiency Interview – Computer as an Alternative

The Oral Proficiency Interview – computer (OPIc) is an internet-delivered version of the OPI created by ACTFL to meet the skyrocketing global demand for proficiency assessment. This test administers a series of questions to which the test taker responds orally while being recorded in order to be evaluated later by a trained rater ("Oral Proficiency," n.d.). Thompson, Cox, and Knapp (2016) performed a study to confirm the reliability of the OPIc as it relates to the OPI and showed that there was significant correlation in the scores of the OPI and the OPIc. However, over 70% of the participants in the test said they preferred the OPI to the OPIc and cited primarily the following three motives: interaction with a human felt more natural, topics were tailored to participant interests, and response time for each question was flexible. The 27 of 132 students who stated they preferred the OPIc cited lower anxiety due to not having a rater, the ability to have questions repeated without feeling judged, and the flexibility of not having someone else guide the discussion. The results of this study primarily indicate a preference by the participants for the OPI. In addition, Malabonga, Kenyon, and Carpenter (2005) expressed concern that the self-assessment feature of the OPIc in which a test taker answers questions and essentially chooses at which level to begin the test can lead to lower scores if the individual is

beginning the test at a higher level than he or she should. ACTFL, in response to these concerns, moved to a more comprehensive "can-do statement" questionnaire as the pre-OPIc self-assessment instrument in order to more accurately define the appropriate starting point for test-takers, as it allows them to more clearly identify their level based on a specific list of things they can or cannot do in the second language (Tigchelaar, 2019).

Though the OPIc was originally created with the intention of providing a less expensive, more accessible, and logistically simpler alternative to the OPI, it is still expensive. At Language Testing International, the current cost of a Portuguese OPI is $150 and the OPIc is $105. The need for a more cost-effective alternative to both the OPI and OPIc, while maintaining the same standards of validity, continues to challenge teachers, raters, and researchers in the foreign language community.

Elicited Imitation as an Alternative to the OPI and OPIc

One relatively inexpensive proposed method of oral proficiency assessment that has shown promise is the Elicited Imitation (EI) test. EI consists of a series of L2 sentences which are either read aloud or recorded and played for the test taker, who then repeats each utterance back as accurately as possible. As the exam progresses, sentences increase in length and complexity.

Studies have consistently shown that for test-takers to be able to produce the utterance after hearing it, there must be some level of message decoding and reproduction. Cox and Davies (2012) describe the limited nature of our working memory (WM) and the necessity of interaction between the long-term memory and WM in order for the speaker to break the utterance down into meaningful units and then reproduce it. The more knowledge L2 speakers have built up in

the long-term memory that can be accessed when reproducing the utterance, the more ease with which they will be able to reconstruct the sentence.

Research continues to bolster the case for EI as a valid instrument for measuring oral proficiency for formative purposes (i.e. Bowden, 2016; Cox & Davies, 2012; Erlam, 2006). The primary question remains as to whether or not EI can serve as a cheaper option to the OPIc as an alternative to the OPI. There are some linguistic advantages to the EI, such as the ability to prompt specific grammatical structures and vocabulary usage from test takers that they might otherwise avoid. Additionally, EI testing, as it is simply a series of solitary phrases, possesses the potential to cover a variety of topics in a short time frame. One of the main reasons, which will not be discussed here as it is beyond the scope of this paper, is the potential for automated scoring for the EI test. However, the primary reason for advancing the use of EI proficiency testing is the simple fact that it is much more financially feasible. As previously mentioned, an official ACTFL OPIc currently costs $105, and an OPI is $150, whereas the cost of an Elicited Imitation test is typically under $30. As research continues to show that there is significant correlation between EI testing and OPI formats, the motivation for organizations to use EI as a viable proxy increases (e.g. Burdis, 2014; Wu & Ortega, 2013).

The Present Study

The present study was conducted in order to accomplish two objectives. The primary objective was to create an elicited imitation test for Portuguese that could serve as a valid substitute for the OPI and OPIc. This process was accomplished in collaboration with Emmersion Learning using processes that have previously been utilized in the creation of EI tests for other languages and will be described in detail in subsequent chapters. The second objective was to gain an understanding of test-taker attitudes and reactions to such a test. One of the

primary challenges associated with EI testing is face validity. Regardless of consistent research results showing the value and potential of EI testing, language learners continue to be skeptical of the testing model. The author of this study conducted a survey of participants after they took both the newly created Portuguese EI and then either the OPI or OPIc depending on their EI results. This survey provided insight into the test-takers' experience which will be key to improving future EI test design and administration. This study will attempt to analyze the feedback of test-takers in order to provide suggestions for future practices and procedures that will result in increased test validity.

This chapter has explained the rationale for this study. Chapter 2 will examine previous research on the evolving definition and importance of oral proficiency and then delve into both the OPI and OPIc as measurements of oral proficiency. This will be followed by an analysis of Elicited Imitation testing with a focus on comparisons with the OPI. Some attention will be paid specifically to face validity as a current and prevalent challenge to EI becoming a more broadly accepted instrument. Chapter 3 will provide a detailed description of the developmental process by which the Portuguese EI test was created, highlighting some of the unique aspects of the Portuguese language that made the process more challenging. Chapter 4 will provide an analysis of the test results as well as an evaluation of the survey responses regarding attitudes and reactions to the Portuguese EI experience. The final chapter will provide a brief summary of the study as well as provide some suggestions for the future of elicited imitation.

CHAPTER 2

Oral Proficiency

As noted in the introduction, oral proficiency has been defined and redefined numerous times over the last several decades and remains somewhat nebulous among researchers and educators. This is problematic because before oral proficiency can be consistently and meaningfully assessed, it is necessary to have a standard, or in this case a definition, against which to evaluate results. As far back as the 1960s scholars, educators, national leaders, and others were well aware of the flawed nature of second language education and assessment in the United States. Since then the demand for increased attention to and development of L2 proficiency has resulted in calls for both national and international unification of effort to define and assess language ability. Once such poignant call for consensus on understanding proficiency came from Omaggio (1983):

> By agreeing on what it means to know a language at various stages of competence, and by describing what a person can typically do with the language at each of these stages, we can begin to find a way to measure outcomes against a common metric, and to predict accurately the degree of success with which an individual can handle a variety of needs in a whole range of situations. The descriptive power that we can obtain from this common metric can help us compare and contrast more intelligently the effects of existing methods and materials. (p. 330)

A key phrase in this quotation that remains at the core of the modern definition of proficiency, particularly oral proficiency, is the concept of identifying what a person can *do* with the language. Chomsky (1965) differentiated between that which individuals know and understand about a language, which he called their "competence," and their verbal output or

behavior, which he called “performance”. It is the difference between being able to identify and describe the uses of a shed full of tools versus the ability to use them in symphony to construct a box or a palace. This idea was central in the 1980s as scholars such as Omaggio were attempting to define the often used term of *communicative competence.* In his article on the topic, Canale (1983) argued that “…the distinction between communicative competence and actual communication remains poorly understood” (p. 5). His argument was that discussions surrounding communicative competence needed to be just as focused on the user’s skill with the target language as it is on their knowledge about the language. Bachman and Savignon (1986) note that the distinguishing characteristic of communicative competence is

> its recognition of the importance of context beyond the sentence to the appropriate use of language. This context includes both the discourse, of which individual sentences are part, and the sociolinguistic situation which governs, largely, the nature of that discourse, in both form and function. (p. 381)

These ideas gained traction throughout the 1980s and 90s. As a result the concept of proficiency began to evolve into something more definitive.

Within the concept of language proficiency are questions of the different measurable components of language production. For example, how important is grammatical accuracy in the idea of proficiency, as argued by Higgs and Clifford (1982)? This includes syntax, morphology, lexicon, and other sentence-level mechanisms. How well do speakers adhere to sociolinguistic rules of appropriateness? Are they able to employ various verbal and nonverbal strategies to compensate for deficiencies in other areas of communicative competence (Bachman & Palmer, 1982; Omaggio, 1983)?

Over a period of a few decades, different organizations around the world seemed to follow the same trend, which was to create a descriptive scale of some sort that gave explanations of what different levels or ranges of oral proficiency should sound like. Some scales were extremely detailed in their descriptions whereas others were much more general. Brindley (1998) provided examples such as the Australian Second Language Proficiency Rating Scale which is particularly descriptive in nature, discussing details such as pronunciation, word order, vocabulary, tense usage, descriptive ability, comprehensibility, and much more. In contrast, Brindley highlights the general nature of the English Speaking Union yardstick scale which focuses more on effectiveness of communication and clarity of the speakers message. The creation of these various scales was shaped in large part by their creators' purposes and intentions for their use. As addressed in the opening chapter, two primary scales were a result of the push for universal assessment parameters here in the USA – the ILR and ACTFL scales. As this project is not a judgement on the validity of the ACTFL Proficiency Guidelines as valid criteria, the focus for this literature review will remain on the oral assessment instrument (the OPI/OPIc). These topics were briefly mentioned in the opening chapter but will be addressed in more detail in the following sections.

OPI and OPIc as Instruments of Oral Proficiency Measurement

The Oral Proficiency Interview

Assessment of oral proficiency is by nature only accomplished in one way. Regardless of the organization, country, model, or scale, in order to evaluate the speaking ability of a second language learner in that target language it is necessary to elicit some variation of a speaking sample from the individual. The conditions and methods by which this is accomplished have varied over the years. The ACTFL OPI is one such instrument that is widely accepted across

industries and borders as a valid tool for oral proficiency assessment. As with any instrument, the detractors and naysayers are not few in number and their arguments are not without merit. Many would contend that ACTFL has not done its due diligence in establishing test validity and that the different level descriptions were created somewhat arbitrarily or that the concept of the "native speaker," as is regularly referenced in the guidelines, is an unacceptably oversimplified personification of the target audience with whom the test-taker would hypothetically be conversing (Bachman & Savignon, 1986; Chalhoub-Deville & Fultcher, 2003; Lantolf & Frawley, 1985). The validity of these opinions is also not the focus of this project. These, among other arguments, have done little to slow the advance of the ACTFL OPI as the primary tool for companies, universities, and government agencies to make personnel, placement, and admission decisions regarding language speaking ability. Therefore, it is important in the context of this project to understand the history and theory behind the ACTFL OPI.

The OPI is essentially as old as the ACTFL Proficiency Guidelines themselves, and the procedure of eliciting ratable speech samples through oral interviews actually predates the Guidelines. The Guidelines and the interview are based on assessment methods originally developed in the 1950s by the U.S. Department of State. In the 1940s it became painfully clear to military and government leadership that key personnel were woefully unprepared to meet the language demands of World War II (Kaulfers, 1944). The Army Specialized Training Program was developed in response to this deficiency in the forces, which resulted in the first real rating scale that consisted of what we might call the original "can-do" statements in language assessment (Chalhoub-Deville & Fulcher, 2008). Kaulfers (1944) published the following four-point scale:

1. Can make known only a few essential wants in set phrases or sentences.

2. Can give and secure the routine information required in independent travel abroad.
3. Can discuss common topics and interests of daily life extemporaneously.
4. Can converse extemporaneously on any topic within the range of his knowledge or experience. (p. 144)

A similar six-point scale was later developed by the Foreign Service Institute (FSI) in response to the Korean War. This scale was adopted by a number of government organizations such as the Central Intelligence Agency, Defense Language Institute, and the Peace Corps. In the late 1960s these agencies came together to form the Interagency Language Roundtable (Chalhoub-Deville & Fulcher, 2008). The ILR scale runs all the way up to 5 as the highest score possible:

- 0 – no proficiency
- 0+ – memorized proficiency
- 1 – elementary proficiency
- 1+ – elementary proficiency, plus
- 2 – limited working proficiency
- 2+ – limited working proficiency, plus
- 3 – general professional proficiency
- 3+ – general professional proficiency, plus
- 4 – advanced professional proficiency
- 4+ – advanced professional proficiency, plus
- 5 – functionally native proficiency (Interagency Language Roundtable, n.d.)

However, the majority of second language learners will likely remain in the 2+ or lower ranges, meaning that there are only six levels beginning at 0 with which to describe most speakers' abilities. However, a score of 0 actually represents no real functional ability in the language, so

in actuality there are only five levels. Researchers and educators were concerned that although this scale might be sufficient for military and government purposes, a scale that would be useful in universities and other classrooms would need more precisely defined levels by which to categorize learners.

In the late 1970s and early 80s the U.S. government elected to spend unprecedented amounts of money on an initiative to create such a program. ACTFL and the Educational Testing Service (ETS) were awarded grants to achieve this goal which resulted in the publishing of the ACTFL Provisional Proficiency Guidelines in 1982, with the publication of the completed guidelines to come later in 1986. Training for the first oral proficiency raters based on the new guidelines began in that same year as the provisional guidelines and continue today (Liskin-Gasparro, 2003). When the training manual for raters was published in 1999, the expectation was that the ACTFL OPI would be utilized in decision making in any number of government or private organizations regarding employment, graduation, admission, promotion, and much more (Swender, 1999). The current Oral Proficiency Interview Familiarization Manual published by ACTFL (2012) states the following:

> The ACTFL Oral Proficiency Interview (OPI) is a valid and reliable testing method that measures how well a person speaks a language…The ACTFL OPI is currently being used for a variety of purposes in academic, commercial, and government contexts. Because an ACTFL OPI rating provides a common metric for describing a speaker's functional ability in a language, it serves as a way of providing articulation among language programs. OPI ratings are commonly used for purposes of admission into programs, placement within a language sequence, and determination of the fulfillment of exit or graduation requirements. Teacher Certification Boards in some states require evidence of

> spoken language competency as demonstrated through an Official ACTFL OPI. In some cases an ACTFL Official OPI rating can be used to waive certain language requirements for teacher certification. The OPI and OPI ratings are widely used in the business world and in government for purposes of hiring and promotion in multilingual positions. (p. 8)

This extensive list of uses for the OPI illustrates the impact it has had over the language acquisition community. The resulting influence of the OPI in professional communities has had sweeping effects on language instruction that has reached all the way down into the elementary school classroom (Liskin-Gasparro, 2003). Despite the ongoing insistence by many in the flaws of the ACTFL OPI process, its utilization continues to increase. The influence that the OPI and the ACTFL guidelines have had on the proficiency movement, while not totally understood nor thoroughly researched and analyzed, played an important role in the "paradigm shift" away from measuring students against absolute grammar and vocabulary norms to the pursuit of communicative objectives (Swaffar, Arens, & Byrnes, 1991).

The OPI Process

Because the elicited imitation test as developed through this project was created with the intention of being used as a reliable and valid substitute specifically for the ACTFL OPI/OPIc, it is important to understand the process and experience of taking the OPI. The ACTFL official website (n.d.) describes the OPI as follows:

> It is a 20-30 minute one-on-one interview between a certified ACTFL tester and an examinee. The interview is interactive and continuously adapts to the interests and abilities of the speaker. The speaker's performance is compared to the criteria outlined in the ACTFL Proficiency Guidelines 2012 - Speaking or the Inter-Agency Language Roundtable Language Skill Level Descriptors – Speaking. The interview is double rated

and an Official ACTFL Oral proficiency Certificate stating the candidate's proficiency level is issued to the candidate.

Although there are many criticisms that can and have been leveled at the ACTFL Proficiency Guideline/OPI system's declaration of validity and reliability, the most pertinent to this study is the issue of raters. Because the OPI is a face-to-face or telephonic approach to oral proficiency that relies on the judgement of a trained rater (Malone & Montee, 2010), it is necessary that each of the raters be trained to replicate a specific process of sample elicitation that varies as minimally as possible from the thousands of other interviews completed by certified raters in nearly 100 languages. The general steps to that interview process as described by Malone and Montee (2010) are outlined below:

- Warm-up: In this first step of the interview the rater is having a simple conversation with the examinee in order to put them at ease and also to gain a basic idea of their conversational level as well as to establish a list of topics mentioned by the interviewee that the rater can use to create questions for level checks and probes (described below). Some simple questions in this initial stage of the interview might include *Where are you from? Where does your family live now? What do you do for work?*
- Level Checks: This portion of the interview is designed to allow the rater an opportunity to discover the level at which an examinee can sustain a dialogue. These checks are very specific, meaning that the questions asked and the conversation will be intentional in the level they are inducing. In other words, a level check for the ACTFL Intermediate level would correlate directly with the tasks described in the guidelines from the Intermediate level.

- Probes: The interviewer will then attempt to discover the level at which the examinee will experience a "breakdown" in conversation, meaning the level at which the learner can no longer sustain the dialogue. These probes will typically be one level higher than the level check. By alternating back and forth between level checks and probes the rater can carefully identify the specific level at which the examinee can consistently sustain conversation.
- Role Play: The interviewer and the examinee take on different roles as described on a role play card written in English that describes the situation to be acted out. The purpose of the role play is to introduce a complicated situation that will serve as either a level check or probe for the rater to achieve further clarity on the learner's ability. An example of an intermediate role play might be *Imagine you awaken feeling sick on the day of an important test. Call your professor's office to explain the situation and decide on a plan of action.*
- Wind Down: The interviewer will ask some questions that are at or below the conversational level of the examinee in order to cool down and end the interview in a positive state such as *Do you have anything else going on today? Any plans for the long weekend?*

This entire conversation is recorded and rated by the interviewer and then passed to an additional rater for their rating in order to increase reliability. Surface and Dierdorff (2003) completed a study of 5881 ACTFL OPIs in 19 different languages in order to analyze interrater reliability. 4751 (80.79%) of the OPIs received the exact same result from both raters, while 1093 (18.59%) were separated by one sub-level, 34 (.58%) by 2 sub-levels, and only 3 (.05%) by 3 sub-levels. A holistic interpretation of these results would support a strong argument for

interrater reliability, notwithstanding the previously mentioned arguments against the validity of the guidelines and OPI as a whole.

Once the dual-rater process is completed and a final score is agreed upon, the results are delivered to the test-taker. This whole process takes several weeks. An individual interested in signing up for an OPI will need to schedule it a couple of weeks in advance and can expect to wait up to two weeks for the results to come back. Because the OPI is a face-to-face or telephonic interview by trained interviewers, the associated costs of training and employing raters can be relatively high.

The OPIc Process

The scheduling and financial challenges associated with the OPI make it an unrealistic option for many interested parties. In an attempt to make the test more affordable and practical, ACTFL created a computerized OPI known as the OPIc (Thompson, Cox, & Knapp, 2016). The OPIc shares many of the same characteristics as the OPI in that it is an interview style test that is recorded and rated by two raters and given a score based on the ACTFL levels and sub-levels. The primary difference is that instead of the interviewer being a live person, it is an avatar on a computer screen. “The goal of the instrument is the same as the OPI: to obtain a ratable sample of speech which a rater can evaluate and compare to the ACTFL or ILR Proficiency Guidelines in order to assign a rating.” (Oral Proficiency Interview by Computer, n.d.)

The other major difference is that there is a background survey and self-assessment. The purpose of the background survey is to identify some topics of discussion related to the test-takers work, school, and personal activities. Because there is no live interviewer that can react to question responses and adapt each test to the individual test-taker, this background survey is necessary. The self-assessment is used to identify an appropriate range of proficiency for the test.

There are five possible test formats resulting from the self-assessment responses based on the different guideline ranges. This allows the interview to cover only the likely result levels instead of spending time on Novice or Superior level questions, for example, if all indications from the self-assessment indicate a likely Intermediate or possibly Advanced result. This keeps the interview within a reasonable time frame.

Though the OPIc is increasing in popularity based on convenience of scheduling and affordability, there are still questions of concurrent validity between the OPI and OPIc. One study completed by Surface, Poncheri, and Bhavsar (2008) compared the scores of participants across the two testing formats and generally speaking their results showed strong comparability between the two tests. The primary concern was the lack of absolute agreement on scores, which occurred only 63% of the time. However, that number rose to 98% if one sublevel of discrepancy was allowed, meaning Intermediate High versus Advanced Low, for example. In response to this study ACTFL made some changes to the OPIc and the same group of researchers completed another study with improved results (Surface, Poncheri, & Bhavsar, 2008). Absolute agreement increased to 87% and 100% when adjacent sublevels were included.

There are still concerns that need to be researched regarding the OPIc, such as the effects of the self-assessment and issues of test-taker preference (Thompson, Cox, & Knapp, 2016). However, for the purposes of this project, the elicited imitation test that was created is intended to be a substitute for either format based on the premise that they are approximately equivalent in validity and reliability.

Elicited Imitation

Elicited imitation is not a new idea and has faced substantial scrutiny in its development and proposal as a valid method for assessing L2 oral proficiency (e.g. Bley-Vroman &

Chaudron, 1994; Vinther, 2002). For decades the second language community has been interested in the concept of elicited imitation and its potential use as tool in proficiency assessment. A typical EI test consists of a number of stimuli that are intended to be played or read for the test-taker who will then repeat them back as precisely as possible. Historically, EI was mainly used with children to determine their understanding of grammatical features (Bley-Vroman & Chaudron, 1994; Gallimore & Tharp, 1981). The assertion made by EI advocates is that in order for a learner to be able to accurately imitate the grammatical structure of a prompt, there must be previous target language understanding and experience with it (Drackert, 2015). Bley-Vroman and Chaudron (1994) explain that learners must have accurately parsed the sentence through their developing grammar and comprehended the meaning. Over the past several decades EI has developed into a tool for overall language proficiency. There are several reasons for this according to Drackert (2015). The first is that it taps into all of the core abilities of language production in a very short period of time, including vocabulary, phonology, and morphosyntax. Second, it eliminates the variance that exists in the OPI and OPIc as a result of the existence of an interviewer, giving the administrator a high degree of control. Another important feature of an EI test that is not discussed in this study is that it does not require any level of literacy from the test-taker and can therefore be administered to adults and children as well as low- or non-literate populations. Finally, and most importantly to this study, is that it offers a low-cost and logistically simple option for those without the resources to access tests like the OPI or OPIc.

The Problem of Face Validity

A test is valid to the extent that it measures what it purports to measure. The term *face validity* refers to the degree that it *appears* to measure what it claims to measure. In other words,

does it look like it does what it says it does? There is very little research that addresses the issue of elicited imitation and face validity specifically. In her doctoral dissertation, Gaillard (2014) mentions the problem of face validity from the perspective of language teachers who are taking a communicative approach to their language instruction. It is understandable that the idea of assessing a student's communicative ability through the repetition of a series of unrelated sentences might be hard to accept as valid. The connection between this test method and communication ability might be hard to make (2014). Van Moere (2012) stated this problem most plainly in the following excerpt:

> The elicited speech is decontextualized rather than situated in a meaningful dialogue, is primed and constrained rather than spontaneously constructed, and moreover is lacking in purposeful communication. For learners, teachers and researchers inducted into communicative methods, or where practice and rehearsal are not commonly utilized, the tasks may appear too divorced from real-life interactions, or too narrow in construct, to yield information about test taker ability that can be generalized to real-world domains. (pp. 339-340)

An additional critique of EI testing is the possibility that participants are actually not understanding the items that are being repeated but that they are simply "parroting" the sounds through the use of WM (Bowden, 2016; Vinther, 2002). Miller (1956), in his discussion of immediate memory span, proposed that the human brain could only maintain in memory 7±2 unrelated items, such as words or sounds in this case. He also gave birth to the idea of recoding, which is more commonly known as *chunking* in language learning discussion. This is the idea that instead of remembering individual words, certain word patterns are "chunked" together in order to minimize the processing work required of the WM. Half a century later, Cowan (2001)

put the magic number at 4±2. This number was confirmed by a study performed to gauge the impact of WM on elicited imitation testing. Okura & Lonsdale (2012) compared EI scores of English learners at the English Language Center (ELC) at Brigham Young University (BYU). The students had already been given an English "level" by the ELC. They were given a test designed to assess WM (unrelated to language or English), which confirmed the 4±2 estimate of Cowan, as well as an English EI test. There was no significant statistical correlation between EI scores and WM scores nor between ELC levels and WM scores. However, there was significant correlation between ELC levels and EI scores, indicating that the role of WM in EI is minimal (Okura & Lonsdale, 2012).

Other studies have yielded evidence that EI tests actually do require comprehension of meaning. One study cites the fact that the absence of context and prior knowledge about what test takers are being asked to produce serves as evidence to the argument that the individual must have an understanding of meaning in order to be able to reproduce the solitary sentence (Erlam, 2006). Erlam, as well as Drackert (2015), referenced cases where participants presented with ungrammatical utterances would spontaneously reconstruct the sentence into a grammatical form, further supporting the claim that it is meaning that is registered in the memory of the participant and not simply a series of words and sounds. This issue will arise again later in this study when addressing the attitudes of the test-takers of the Portuguese EI test.

EI as a Substitute for the OPI/OPIc

The primary concern addressed by this study as it relates to the validity of EI testing is whether or not it is a valid replacement for the OPIc. In pursuit of that answer I will assume that the latter is a valid proficiency assessment tool, notwithstanding the previously mentioned criticisms of its validity. Similar studies comparing EI tests with the OPI/OPIc have been carried

out in other languages (e.g. Burdis, 2014; Gaillard, 2014; Wu & Ortega, 2013). As far back as the early 1980s researchers were beginning to compare elicited imitation with interview-style tests. At a time when the OPI was still in development, Henning (1983) compared the results of an EI type test with a Foreign Services Institute interview as well as a sentence completion section in which participants were given a three word (or less) introductory piece of a sentence that is to be completed by the participant. This study involved 143 Egyptian students learning English. In what were considered surprising results, the highest overall validity was found for the imitation method. In a study performed by Cook, McGhee, and Lonsdale (2011), 25 ESL students of varying backgrounds took an English EI test in order to discover how accurately their score would predict their scores on an OPI. After several iterations of the experiment using a variety of word corpora as well as a number of subsets of test items the correlation between predicted OPI scores and actual OPI scores was found to be over r =.80 in all but 7 of the 59 subsets. Lonsdale and Lever (2015) completed a similar study in which 42 Portuguese learners were administered both the OPI and an 84 item EI test. The tests were graded twice by humans with the results showing very strong correlation with r = .93 in the first round of grading and .92 in round two.

The comparability of EI tests and the OPIc has also been addressed previously in studies of learners of Spanish as well as other languages. In a study of Spanish college students comparing the Center for Applied Linguistics' Simulated OPI (SOPI), which is based on the ACTFL scale, with EI, the results showed significant correlation between student SOPI and EI scores with r = .88 (Ortega, Iwashita, Norris, & Rabie, 2002). Finally, in a study of 37 native English speakers learning Spanish, Bowden (2016) found that there was a very high correlation

(*r* = .91*)* between the SOPI and EI test up to but not including the very highest levels of the ACTFL scale (Superior or Distinguished).

Though previous Portuguese EI tests have been created, such as the abovementioned Lonsdale and Lever analysis, no studies have been completed that compare them directly to the ACTFL OPIc. One of the purposes of this project was to address this gap in EI research and testing. In addition to validating the test, there is also a requirement for a better understanding of the effect of test-taker attitudes and face validity on the general acceptance of EI as a valid tool for oral proficiency measurement. This study aims to provide additional information about these issues.

CHAPTER 3

Development and Validation of the Test

This chapter will provide the details of the process used in collaboration with Emmersion Learning (EL) to create the Portuguese Elicited Imitation Test. This process is one that has been previously used by the company to create EI tests in other languages but possesses some unique aspects for the Portuguese language that will be covered in detail.

Building the Corpus

The first step to creating the test was to creating a set of prompts that would accurately identify the separation of different proficiency levels. The final EI test that would be commercialized by Emmersion would be a 30-item test, but in order to arrive at that point we first created a 60-item calibration test that was administered to the participants in an effort to identify the best 30 items. Because the standard by which the EI test was attempting to establish concurrent validity was the ACTFL OPI and OPIc, the development of an item corpus was also centered on identifying prompts that would reflect the speaking abilities of language learners at the four main levels: novice, intermediate, advanced, and superior. To this end, it was necessary to create a range of sentences that would represent the different levels. This required the involvement of native speakers. Ten native Portuguese speakers were given a test that began with prompts that were 15 syllables long and then increased in length until a capacity limit was discovered. That length of prompt, based on syllable count and not word number, was set as the "superior" threshold from which the other levels would be established. Using the previously created Spanish EI test as a guideline, the initial syllable counts established for each level were as follows: superior 25-35; advanced 18-28; intermediate 12-22; novice 5-14. More explanation of levels will follow in subsequent sections.

All of the original sentences identified by EL as potential prompts for the test were found on Tatoeba.org, which is simply a collaborative website containing millions of sentences translated into hundreds of different languages. Contribution is free and open to all. The user can simply select a primary language and a target language, type in a word or phrase in the primary language, and a series of sentences that include that word or phrase in both languages will be displayed. However, because there is no restriction on who can contribute sentences and their translations, every result is suspect. An example of a search would look like this:

Translate from: English *Translate to:* Portuguese

Word or phrase to translate: chicken soup

Result: I would like chicken soup.

Translation: 1. Eu gostaria de sopa de frango. 2. Quero canja de galinha.

Result: Your chicken soup is great.

Translation: 1. A sua sopa de galinha é ótima. 2. A sua canja é ótima.

Result: No one makes chicken soup like my mother.

Translation: Ninguém faz sopa de frango como a minha mãe.

This example shows the potential for a variety of resulting translations as well as the increasing length of the results. If the user were to search a single word the search would likely return hundreds of results ranging in length from two or three words to more than twenty.

In the initial gathering process, EL created at random a corpus of one thousand sentences of varying lengths. These sentences were delivered to myself and one other Portuguese expert whose identity is not known to me in order for us to identify fifty to seventy prompts for each of the ACTFL levels which could serve as potential prompts in the final EI test. I will discuss my

methods for item elimination and selection in the following paragraphs. The methods of the other expert remain unknown to me.

The primary factor in making these decisions was the length of the prompt. Past research in EI testing shows that the majority of the variance in difficulty of a prompt was related to its length. In their analysis of a Chinese EI test Wu and Ortega (2013) discovered that there was a significant negative correlation between the average score of a given item and its length. Their data showed that 74% ($r_2 = .74$) of the variance was attributable to sentence length. In another study of 81 English learners, researchers came to a similar result that 73% ($r_2 = .73$) of difficulty being based on sentence length (Graham, McGhee, & Millard, 2010). However, there were several other factors that were taken into consideration when determining the potential use of a sentence. In these same studies mentioned above, other factors that contributed to item difficulty variation were lexical frequency, lexical density, and morphological complexity. On a more basic level, it was first important that we eliminate anything with proper nouns, triggering or offensive content, or terms adopted from other languages.

Non-Grammatical Elimination

As stated previously, I relied primarily on sentence length to give a base from which to continue separating sentences into the most appropriate ACTFL levels. My initial scan of the list of potential sentences was a simple search for sentences that had obvious ambiguities in meaning or unnatural structure. One such example was the following sentence:

Eu sou casada, você é casado, vamos nos casar!

(I am married, you are married, let's get married!)

Other sentences were removed for posing potentially inappropriate or offensive interpretations, such as this sentence:

Quero que você morra.

(I want you to die.)

There were other reasons that did not fall into lexical or morphological categories for which sentences were removed. In some cases, there were sentences that were very similar to each other or several sentences that contained many of the same words. For example, there were well over 100 sentences that addressed the topic of learning to speak French. One of the two sentences below was removed due to its repetitive nature.

Eu estou procurando uma das minhas irmãs.

(I am looking for one of my sisters.)

Eu estou procurando um dos meus irmãos.

(I am looking for one of my brothers.)

I also attempted to remove any sentence that might force a female test-taker, when repeating the prompt, to refer to themselves as a member of the opposite sex, or vice versa. This was not done for reasons of propriety or sensitivity, though these might be considered valid reasons for their removal. My motivation for doing so was simply to avoid forcing a participant to do something that was so unnatural that it might cause them to subconsciously make a mistake. The following example illustrates this problem:

Eu sou a melhor amiga dela.

(I am her best friend.)

A male speaker repeating this sentence, in order to receive full credit, would be forced to repeat this sentence verbatim and would therefore have to refer to himself as a "female" friend. A natural tendency in this case would be to replace the word *amiga* with *amigo*, or "male" friend,

as well as replacing the article *a* with *o*, resulting in two potentially biased errors in what should be a simple prompt.

The last non-grammatical reason for removing a prompt worth mentioning were those that could be interpreted as containing two voices or speakers, such as the following example:

Como ocorreu o acidente? Ninguém sabe.

(How did the accident happen? Nobody knows.)

Generally speaking, I did not want to include any prompts that contained more than one sentence in order to avoid the possibility of students mistakenly believing they had already heard the whole prompt when there was still more coming. However, in some cases I saved multi-sentence prompts at the 30-35 syllable level.

Grammatical Elimination

There were a number of grammatical reasons for removing sentences as well. The first of this group were the sentences that contained grammatical structures that are not used in both Brazilian and Continental Portuguese. Some examples are the following:

Dar-te-ei uma maçã.

(Give you, I will, an apple.)

O carro está a esperar no portão.

(The car is waiting at the gate.)

Meu avô costumava fazer os próprios móveis.

(My grandfather had the custom of making his own furniture.)

The first example shows a verb/indirect object structure that would rarely be used in Brazilian Portuguese. This is not to say that it is incorrect in Brazil, but it is an overly formal structure that would be unfamiliar to a novice or even intermediate learner. The second example contains one

of the most glaring differences between Brazilian and Continental Portuguese, which is the use of the gerund. Brazilian Portuguese uses the gerund just like English, where the suffix *-ndo* is a direct translation of the English suffix *-ing.* In Portugal a different grammatical structure replaces the gerund with the preposition *a* followed by the infinitive form of the verb, as seen in the example above (*a esperar* = *waiting*). This construct was featured in many of the proposed sentences. This may appear to be a bias towards Brazilian Portuguese, but the reason for this step in the elimination process is that while the gerund is used and understood by speakers of Continental Portuguese, the reverse is not always true of the *a + infinitive* structure amongst Brazilian Portuguese speakers. The third example above illustrates another common difference in the two dialects. In Continental Portuguese it is nearly the rule to use an article prior to a possessive pronoun. In the example above the sentence begins with *Meu avô* (*My grandfather*). While this would be an acceptable structure in Portugal, a test-taker accustomed to Continental Portuguese may subconsciously add the article to the beginning of the sentence (*O meu avô)*, resulting in an extra syllable in the examinee's response. Sentences that already possessed the Continental Portuguese structure were acceptable since this is commonly used in Brazilian Portuguese as well.

Lexical Elimination

In addition to grammatical differences, there were also several prompts in which vocabulary had different meanings in the two different dialects. I eliminated these as well to avoid any potential confusion. However, the primary vocabulary-related reason for eliminating prompts was to attempt to correlate length to difficulty. The following examples illustrate this problem:

Ela o apunhalou.

(She stabbed him.)

Alegro-me em vê-lo.

(I am happy to see you.)

As placas tectônicas são pedaços da litosfera.

(The tectonic plates are pieces of lithosphere.)

Diferentes códãos podem codificar o mesmo aminoácido.

(Different codons can code the same amino acid.)

As previously mentioned, research has shown that the vast majority of variance in prompt difficulty is associated with sentence length. Therefore, prompts were chosen largely based on syllable count. This being the case, the sentences categorized as novice level will naturally be shorter than those at the intermediate level and so on. This also means that shorter sentences will primarily contain higher frequency vocabulary. Lexical difficulty will increase as the test progresses through the different levels. The initial syllable counts established for each level were as follows: superior 25-35; advanced 18-28; intermediate 12-22; novice 5-14. Observe that there is an overlap in syllable counts in order to account for the different aspects of difficulty variation. For example, a sentence may only have 12 syllables but contain morphological or lexical aspects that qualify it for the intermediate level. The first example above is only seven syllables, which makes it a novice length sentence, but contains the verb *apunhalar*, which is a low frequency verb and very unlikely to be familiar to a novice or even intermediate level test-taker, and therefore not a good choice as a prompt. The second example is also seven syllables but contains advanced verb structures that would almost certainly be unfamiliar to a novice speaker. The other two sentences have 16 and 21 syllables, respectively, which correlate with intermediate and advanced level. Their lexical content, however, would be considered professional level

knowledge associated with very specific fields and very low frequency. These sentences would be more appropriate for a superior speaker.

Fusion Elimination

The remaining eliminations were made primarily by judging the appropriateness of the relationship between sentence content and length. Many of them, like the samples above, were obvious candidates for elimination. Those that were not as obvious were subject to my judgement of their content. I followed the guidance of Graham, McGhee, and Millard (2008) and attempted to include an assortment of subjects or themes in each level. Another important tool in the elimination process was the Frequency Dictionary of Portuguese by Davies and Preto-Bay (2007). This book was particularly helpful when I was unsure of the frequency of a certain word. Other more "scientific" processes of item selection have been used in other studies, including that of the Graham study mentioned above. However, due to time and resource limitations as well as previous success in EI test creation by Emmersion Learning in other languages using this model, this process was the most appropriate for our situation.

After completing this elimination process, I had selected 75 novice, 77 intermediate, 42 advanced, and 34 superior candidates for possible use in the test. The next step, which actually led to more eliminations, was the syllable count. Accomplishing this task was much more painstaking than one might imagine. It involved two different syllable counts. The first count was a simple summation of the syllables in each individual word to get an initial total syllable count. The second count involved the audio recordings of each prompt, which were also free and available to the public on the previously mentioned website Tatoeba.com. The sentences, as far as can be observed, were all recorded by native Portuguese speakers.

Using the audio recordings, I performed a second syllable count, taking into consideration any combining of syllables, a phenomenon that occurs often in spoken Portuguese. There are a number of situations in which this might occur. Depending on how it occurs, there is a variety of descriptor words associated with each unique scenario. Some of these are vowel fusion, diphthong and triphthong, elision, syneresis, or sinalepha. Determining which of these is the correct term will depend on whether or not the syllables involved are tonic or atonic, the vowels are high or low, vowels are homologous or heterologous, amongst other factors (Fails & Clegg, 2020). As far as this paper is concerned, a full explanation of the different phenomena is not necessary nor realistic. The case of two or more syllables combining to become one is the basic concept that had an impact on this project and will heretofore be referred to as fusion.

In many of the recordings, the difference in the simple count versus the adjusted syllable count, taking into consideration any fusion, was one or two syllables. In some cases, the fusion could be a result of differences of pronunciation or the way in which test-takers separated words. Unlike Spanish, one contributing factor to the more prominent existence of fusion in Portuguese is the fact that indefinite articles in Portuguese contain no consonant, nor do the participle *a (to)* or conjunction *ou (or).* As a result, word final and word initial vowels often combine with these items, which technically constitute their own syllable, to form a single syllable. If this situation is followed by a word that contains the same vowel, the result is a series of words that in their written form contain three syllables but when spoken may be produced as one syllable. As an example:

Seu comentário apoia a ação dos bombeiros.

(His comment supports the action of the firemen.)

The simple syllable separation of the sentence would look like this:

Seu-co-men-tá-rio-a-poi-a-a-a-ção-dos-bom-bei-ros. (15 syllables)

However, taking fusion into account, the syllable separation would appear this way:

*Seu-co-men-tá-rio**a**-poi-**aaa**-ção-dos-bom-bei-ros. (11 syllables)*

The syllables in bold illustrate the effect of fusion. The first instance shows a situation in which the last vowel of a noun is followed by a verb beginning with a vowel. This is extremely common and would be impossible to eliminate from the test entirely. Nor would that be desirable since understanding fusion is an important aspect of Portuguese listening and oral proficiency. The second instance demonstrates a rarer situation in which the final vowel of a verb followed by the article *a* and then a noun beginning with an *a* combined three original syllables into one. These two instances of elision decreased the overall count from 15 to 11. As previously explained, sentence length is the primary driver for difficulty differentiation. More importantly, however, a correct repetition of this prompt may not always sound the same, depending on the way the speaker chooses to separate the words in the sentence. Though a live grader would be able to identify a correct or incorrect utterance by the test-taker, an automated grading system based on voice recognition technology, which is the ultimate objective for Emmersion and their EI tests, might be unable to correctly identify the syllable separation of participants' responses. This resulted in the elimination of several more prompts.

The final syllable count was completed with the remaining sentences. The adjusted count which accounted for fusion was the final syllable count used to divide the prompts into ACTFL levels. As explained previously, sentences with syllable counts of 12-14, 18-22, or 25-28 syllables (overlapping level ranges) were placed into levels based more on lexical and morphological complexity. These groups of sentences were delivered to Emmersion, where they were compared with the recommendations of the other contributor. Ideally, there would be 15

sentences from each level which myself and the other expert had identified as viable candidates for the test. If this were not the case, priority was given to the list created by the academic expert (me in this case). A final collection of 60 items were selected as a result of this process.

Once we had selected the items that would serve as the preliminary test, I located three native Portuguese speakers to record the 60 prompts. I already had audio recordings of the items that had been pulled from Tatoeba.com, but in order to ensure consistency in volume and rate of speech I had them recorded again in a studio. The first recording of the final 60 calibration sentences was done by a native Portuguese female and fellow master's student here at BYU. The others were one male and one female native Brazilian studying English at the BYU English Language Center.

Participant Selection

Throughout the item selection process, I was also recruiting participants to take the item calibration test. The majority of them were students at BYU with varying backgrounds in the Portuguese language. Many of this group were my own students from the classes I had taught over the previous two semesters. These classes were beginning Portuguese for Spanish speakers. Most of the participants from this group had learned Spanish as religious missionaries for The Church of Jesus Christ of Latter-day Saints. A few of them were heritage speakers with only two participants being native Spanish-speakers whose second language was English. This became a question for later as I was interested to know if participants with experience in Spanish would have an advantage when taking an EI test. Emmersion Learning also worked with the Missionary Training Center (MTC) of The Church of Jesus Christ of Latter-day Saints here in Provo to have 31 missionaries that were studying Portuguese in preparation for 18 month to two year service missions to Portuguese-speaking nations. I also recruited several friends and acquaintances of

my own. Our final count was 146 participants from all different levels of the ACTFL scale. Motivation for participation was not something I addressed in this study, but certainly the chance at a free ACTFL Certificate for the OPIc was a motivator for many, something that would cost them over $100 otherwise. The missionaries were obligated to participate as trainees at the MTC, and many others participated as a result of our personal relationship.

Test Administration

Participants were sent an invitation by email that allowed them to select a time for their test and also took them through a brief survey related to their previous foreign language experience. The first questions asked about any family history with the Portuguese language. Then they were allowed to assess their own fluency. The reason for this self-assessment as a separate requirement from the self-assessment they would receive as a part of the OPIc was to ensure that we had recruited a sufficient number of participants at each ACTFL level. Though this survey sent out by EL was not designed to identify specific levels, it did give us an idea of what levels were lacking in participants. Finally, they were asked to do a self-assessment of their Spanish speaking abilities so that we could assess any connection between Spanish experience and success on a Portuguese EI test.

The test was administered in a computer lab at BYU. Students were given an explanation by the test administrator from EL of the process of the EI test. Participants used headsets with microphones to complete the EI test. After completing the 60 item EI test, they went through the OPIc background survey as well as the self-assessment as explained previously and then completed the OPIc. After the test, several participants complained that they were unable to focus very well due the voices of other participants around them. Unfortunately, the test

administrator did not note the names of these participants and we were unable to assess any relationship between their scores and their complaints about the testing environment.

Item Rating and Validation

Once the participants had all completed the test, I and one other Portuguese master's student at BYU were trained to complete the manual rating portion of the analysis process. This process was simple and straightforward. Emmersion had developed a program in which the raters could listen to the recording and simply mark each syllable as having been accurately pronounced in the response or not. We also had access to the original native recordings to help clarify any uncertainties.

Using the scores of the individual items for each participant, the psychometrician at Emmersion was able to identify any test items as well as any test records that were misfitting and remove them from the data set. This was accomplished by calculating infit and outfit statistics of items and response patterns (TrueNorth, 2019). Rasch reliability statistics were then used to determine if the test items were individually sensitive enough to identify low- and high-level speakers. In order to examine the reliability between the examinee's estimated ability and their ACTFL levels as would be determined by the OPIc portion of the test, the psychometrician employed Spearman's rho (TrueNorth, 2019).

Another important objective of the analysis was to determine whether or not the test was biased in any way toward those who speak Spanish as a result of the similarities between the two languages. As many of the participants in the calibration test were either returned missionaries from Spanish-speaking missions with extensive Spanish experience or heritage Spanish speakers, it was not only an ideal opportunity to make this determination, but it was also important to

ensure that the data was not skewed in any way as a result of this particularly unique demographic of Portuguese learners.

Finally, the OPIc scores were used to provide concurrent validity to the EI test as well as to create the scoring algorithm for predicting future EI test-takers' ACTFL proficiency levels (TrueNorth, 2019). One of the enticing features of the EI test that Emmersion markets is its ability to provide the test-taker an accurate prediction of what they would likely score on and ACTFL or Common European Framework (CEFR) OPI. These results will be discussed in the following chapter.

Post-Test Survey

After the tests had all been administered the original plan was to then complete a series of focus groups with several of the participants in order to better understand attitudes and experiences with the EI test. The objective of these focus groups was to better understand the aspect of face validity as it pertains to the EI format and how this issue might be addressed. However, due to the arrival of summer and students returning home, the unavailability of the missionaries, and several other unforeseen circumstances it became logistically impractical to organize the focus groups. Since understanding the experience and attitudes of the participants was one of my primary objectives, I opted to create a survey that could accomplish this goal. Upon further consideration I believe that this may end up being the better option as I am concerned that my presence in the focus groups may have influenced responses since many of the participants were either my personal acquaintances or previous students. It is also important to note that three months had passed between the time of the test and the distribution of the survey. This gap likely resulted in a lower response rate.

Of the 146 participants who took the test, 39 responded to the survey, which is a 27% response rate. Several of the participants emailed me to say that they would know better how to reply to the survey questions if they could first get the results for the EI portion of the test. However, because the purpose of this first test was to validate the potential of the individual items for use in the final test, there was no overall score given on the EI portion. The results of the survey will be discussed in the following chapter. A list of all of the questions can be found in Appendix A.

CHAPTER 4

Analyses and Results

This chapter will provide a brief review of the statistical results of the EI calibration test in order to establish a foundation of validity from which to address the challenge of face validity through an analysis of the responses to the attitudes and experiences survey.

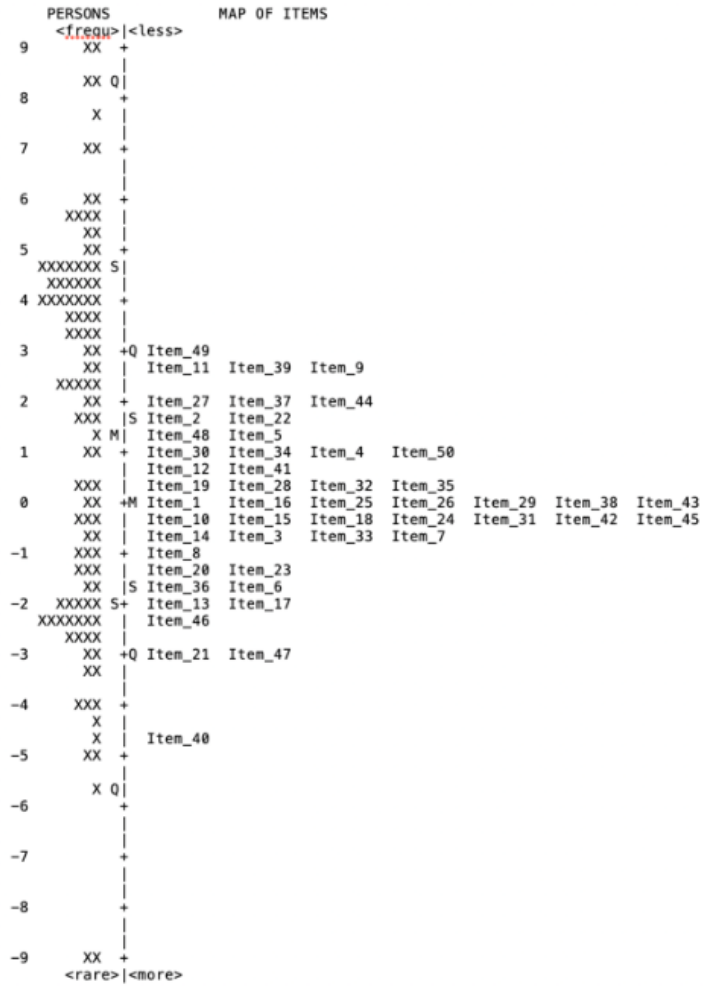
Individual Item Analysis

The purpose of the calibration test was to analyze the individual 60 items identified as potential prompts for the final EI test for validity as well as computing the correlation between the test-takers' predicted ACTFL scores based on the EI test with the actual scores they received from the ACTFL OPIc. To accomplish the first goal several steps were taken. The first step was to eliminate any outliers. By computing and analyzing the infit and outfit statistics of the items and the examinees, 36 test records and 10 test items were identified as misfitting and were removed, resulting in a 110 person/50 item dataset (TrueNorth, 2019). In other words, 36 of 146 participant tests were found to be outliers and 10 of the 60 test items were identified as being poor measurers of the ACTFL skill level they were identified to measure. These tests and items were therefore not used in the calibration process.

In order to confirm an appropriate spread of ability in examinees, Rasch person reliability was calculated at 0.98, indicating that examinees' scores were actually indicative of their abilities. A person separation index was calculated in order to estimate the number of distinguishable ability strata and showed approximately 12 ability strata within the complete range of participants' abilities. This tells us that we have the appropriate spread of ability necessary to accurately identify and separate the different speaking proficiency levels. Rasch item reliability similarly calculated at 0.99, indicating that item difficulty levels were also

reliable. The item separation index, which is used to measure how many separate and identifiable difficulty strata are within the range of item difficulties, indicated that there were approximately 11 difficulty strata that were identifiable. Like the person separation index, this tells us that the spread of item difficulty is sufficient to assess proficiencies at all of the desired speaking proficiency levels.

The following Wright Map shows the distribution of examinee ability on the left relative to item difficulty distribution of the 50 items on the right and indicates appropriate distribution of both item difficulty and examinee ability across the range of abilities (TrueNorth, 2019).



(TrueNorth, 2019)

ACTFL Level Prediction Accuracy

In order to assess the accuracy of the ACTFL predictive ability of the EI test, it was necessary to identify a correlation between the EI predicted scores and the ACTFL OPIc. Applying an ordinal value to the ACTFL sublevels beginning with Novice Low and going up to Superior there are 10 levels. Each level was therefore given a value (Novice Low – 1, Novice Mid – 2…Intermediate High – 6, Advanced Low – 7…Superior – 10). Unlike the ACTFL levels, the predicted scores of the EI test were continuous and could therefore not be directly tied to a score on the ACTFL scale. The EI scores were thus rounded up or down to the nearest whole number and this score was compared to the ACTFL score to see if they fell within one level above or below the actual ACTFL nominal score. It is important to note that ACTFL levels are not separated by equal intervals. For example, the language development required to progress from a Novice High speaker to an Intermediate Low (adjacent sublevels) is much less significant than the development required to progress from Advanced Mid to Advanced High. Because of this, a non-parametric Spearman's correlation was used to calculate the strength of the correlation between the two tests. This analysis showed that there was 72% of the predictions were within one sublevel of actual OPIc scores. Accepting two sublevels of separation increased the level of agreement to 97% (TrueNorth, 2019). As an additional analysis of test validity, the two tests were treated as separate raters of proficiency and then a Spearman rank correlation was used between the two to establish interrater reliability. This analysis showed an interrater reliability of .909 and was sufficiently high to suggest significant correlation (TrueNorth, 2019).

Bias Toward Spanish Speakers

Another objective of the analysis was to identify any bias in the test toward those who self-identified as fluent in Spanish. Since the two languages are similar in so many ways, I was

concerned that because the format of this test is based on a participant's ability to imitate the prompts, there might be an advantage to those who have extensive experience in a similar language. Again, Spanish-speaking participants in the study mostly consisted of those who had served missions for The Church of Jesus Christ of Latter-day Saints in Spanish-speaking regions and learned Spanish as a second language. In order to estimate the influence of speaking Spanish on Portuguese EI scores, the EI test was examined for differential item functioning (DIF) and differential test functioning (DTF), which are measurements of the extent to which an item or test might be measuring different abilities for different subgroups. Analysis of the few items that did show some DIF did not rise to a level of statistical significance. The DTF, which tested the average directional bias, also indicated that one test did not favor one group over another (TrueNorth, 2019).

EI Experience Survey

The preceding data provide a compelling case for the validity of the Emmersion Learning Portuguese Elicited test. Having completed the validation process in several other languages previously, this outcome was largely expected by EL and myself. The data from this project further solidify the many previously made arguments for the validity of EI testing as a substitute for the OPI process. Establishing validity was important to provide a basis from which to better understand some of the challenges facing EI testing and its adoption into the academic, professional, and government realms as a valid tool for foreign language oral proficiency assessment. One of the constant impediments to the acceptance of the EI is the question of face validity. The following paragraphs will provide an analysis of the results of the survey that was provided to participants after they had completed the EI calibration test and the OPIc and was outlined in the previous chapter.

The survey was designed to give participants an opportunity to express their reactions to the experience of taking and EI test. Most of them were completely unfamiliar with this testing format and this showed in their responses. The survey also asked them to make comparisons between the OPIc and EI experiences. In some cases, I have attempted to categorize the participants' responses into themes in order to analyze and understand the overarching concerns or opinions regarding the elicited imitation experience. The objective of this analysis is to be able to propose some methods for minimizing these concerns in order to increase the acceptance and use of EI.

Several themes and trends were identified in the 39 responses. The first question was regarding participants' judgment of EI and the OPIc as valid tools of language assessment:

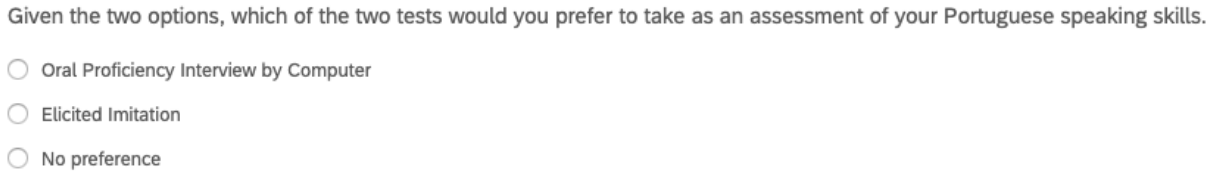
Please answer the following questions comparing the EI test and the OPIc.

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
The elicited imitation test was an accurate assessment of my Portuguese speaking ability.	○	○	○	○	○	○	○
The Computerized Oral Proficiency Interview was an accurate assessment of my Portuguese speaking ability.	○	○	○	○	○	○	○

The first observation I made was that very few of the participants "strongly agreed" with the statements regarding either test. EI had only 2 respondents (5.1%) and the OPIc had only 3 (7.7%) that strongly agreed that these two tests were accurate assessment of their language speaking ability. However, 19 (48.7%) respondents indicated that they "agreed" with the statement regarding the OPIc, whereas only 11 (28.2%) "agreed" that EI was an accurate assessment. However, if all options on the "agree" side are included (somewhat agree, agree, strongly agree), the difference between the two is not so disparate, with EI and OPIc receiving 23 (59%) and 26 (66.7%) responses respectively that fell somewhere in the agreement range. Only 8 participants (20.5%) responded somewhere in the disagree range (somewhat disagree, disagree, strongly disagree) with the statement regarding the OPIc, whereas 15 (38.5%) responded

somewhere in the disagreement range regarding the EI test. Stated another way, approximately two out of every five participants indicated that they felt the EI was not an accurate assessment of their speaking ability, as compared with one out five participants who felt the same way about the OPIc.

The second question gauged participants' personal testing preference:



Five respondents said that they had no preference between the two testing formats. Approximately half of the students (51.3%) selected the OPIc as their test of preference, with the remainder (35.9%) choosing EI.

The following question asked participants to explain why they had selected EI or OPIc or neither. Responses varied but there were some interesting themes noted. In order to accurately make comparisons and draw insight from the responses to this question, I attempted to identify themes that were repeated in more than one response. I identified the following themes:

- EI lacks face validity
- OPIc allows for thought and creation
- The OPIc is difficult to understand
- EI is a test of memorization
- EI has no requirement for understanding meaning

It should be noted that participants did not use the words "EI lacks face validity" nor any of the other exact phrases I have identified as themes, but that these were my interpretations of the ideas they were conveying in their responses. By far the most common themes were the claim

that EI lacks face validity (8 responses) and that the OPIc allows for thought and creation (9 responses), implying that EI does not. Three of the responses that claimed in some way that the EI lacks face validity also specifically mentioned that EI was simply a test of memorization. I identified this as a specific theme so as to distinguish this face validity concept from others. Several of the responses that highlighted a lack of face validity also touched on the idea that EI does not require the test taker to understand meaning. However, three of the participants recognized the need for comprehension in order to be successful on EI tasks.

There were also several complaints that the tests were difficult to understand. Whether the examinees were referring to an audio issue, accent challenges, or vocabulary difficulties is unclear. None of the responses to the question about technical difficulties mentioned any issues with the audio. I therefore have to assume that students were referring to a difficulty understanding the test items for another reason, such as speaking style, context, or vocabulary.

In a follow-up question the participants were asked to share any other thoughts regarding either test method. One interesting theme that emerged in response to this question was the number of respondents that mentioned how discouraged they felt at the end of the EI test. Some of these respondents actually mentioned that they were learning the language in the MTC at the time and were therefore most likely somewhere in the Novice range, and that an experience such as this one merely highlighted their deficiencies. Some examples of participants' comments follow:

- *I believe it would be more productive to measure comprehension in a way that doesn't result in a universal lack of effort.*
- *I know it's just to see how we're doing with the language but I felt like I would never learn Portuguese after I took those.*

- *…it sounded very robotic and made me feel pretty sad about my Portuguese after 6 weeks in the MTC.*
- *The difficulty curve seemed randomly timed, and there was no point after the difficulty where it went to an easier point to restore confidence.*

Another important theme that came up in at least three of the responses was that the testing environment was not ideal. Administering the test in a computer lab to several participants simultaneously resulted in distracting ambient noise.

The next question dealt with stress levels associated with the individual tests. Results showed minimal difference between the two tests:

	Strongly disagree	Disagree	Somewhat disagree	Neither agree or disagree	Somewhat agree	Agree	Strongly agree
The instructions for taking the test were easy to understand.	○	○	○	○	○	○	○
Taking the OPIc was a stressful experience.	○	○	○	○	○	○	○
Taking the EI Test was a stressful experience.	○	○	○	○	○	○	○

For no immediately apparent reason, only 35 of the 39 survey respondents answered this question, with 24 (68.6%) respondents indicating that the OPIc test fell somewhere in the stressful range for them (somewhat agree, agree, strongly agree) and 25 (71.4%) responding similarly for the EI test, though only 8 and 9 of these, respectively, were in the "strongly agree" category. Of those that felt that the tests were *not* stressful, only 6 (17.1%) felt that the OPIc fell within the range of disagreement and 8 (22.9%) felt this way about the EI test, meaning that at least 8 people did not feel that the EI test was a stressful experience and at least 6 felt that the OPIc was not a stressful experience.

The final question in the survey asked examinees to provide any suggestions for using an EI test in the future as a tool for foreign language assessment. The results to this question will be

analyzed in the final chapter as a part of the discussion regarding the future of EI and suggestions for further research.

CHAPTER 5

Summary and Conclusions

In this final chapter I will briefly summarize the development and validation of the Emmersion Learning Portuguese Elicited Imitation Test. After discussing the results from both the calibration test as well as the experience surveys, I will draw some conclusions from the data. Finally, limitations to the test will be presented as well as suggestions for future research.

Development of the Test

This particular EI test was developed under the direction and coordination with Emmersion Learning in an effort to create a Portuguese oral proficiency assessment tool that could be used as a more affordable, logistically simpler, and valid substitute for the ACTFL OPIc. The end goal was to identify 30 EI test items that would accurately differentiate Portuguese learners of different ACTFL without the use of an ACTFL Oral Proficiency Interview. The initial task of creating a calibration test of 60 test items was accomplished by gathering a corpus of 1,000 sentences of varying lengths (syllable counts) and slowly eliminating items one at a time based on certain characteristics such as lexical complexity, morphological sophistication, and syllable fusion. I and one other unidentified Portuguese expert were tasked with this process. Initially, our task was to each identify at least 30 potential sentences for each level (Novice, Intermediate, Advanced, Superior). Once this task was complete, our lists were compared to one another in hopes that at least 15 sentences from each of the four levels would be found in both lists. The 60 sentences that were identified from this process were used to make the calibration test that would be administered to the 146 participants.

The test was administered in a computer lab on the BYU campus. Students took the EI calibration test in conjunction with an official OPIc exam. Because the purpose of the EI

calibration test was to assess its ability to predict participants' scores on the OPIc, it was necessary that they take both tests. After all participants had taken both tests the data were delivered to me and one other rater who were trained to identify correctly pronounced syllables in each recording. All 146 tests were rated, one syllable at a time. These ratings together with the OPIc scores were given to a professional statistician in order to statistically evaluate each individual item and its reliability in accurately identifying a speaker of a certain level. In other words, if a participant's overall ability was identified as being Novice High on the ACTFL scale, based on the results from the OPIc as well as the combined indicators from the EI test, we wanted to know how accurately each item on the test reflected that assessment. Could an item that was projected to be an Intermediate High item be answered correctly by a Novice Mid speaker? If so, it was unlikely to be a good indicator of true oral proficiency. Through this statistical process, the final 30 items were identified.

Results of the statistical analysis also revealed that the EI test was not statistically biased toward Spanish speakers, as this was one of my original concerns for EI tests of similar languages. Finally, EI test scores were compared to OPIc scores in order to establish concurrent validity. 72% of EI predictive scores were within one sublevel of actual OPIc scores and 97% were within two, revealing that the EI test was significantly accurate in predicting OPIc scores.

The last component of this process was a survey that was administered to participants in order to understand their attitudes and experiences related to the OPIc and more specifically the EI test. This survey contained questions regarding the participants' opinions about test validity, accuracy, and future use. Of the 146 participants, 39 responded to the survey. The results of these questions were revealed in the previous chapter and will be discussed and interpreted in the following paragraphs.

Discussion of Survey Responses

The Question of Face Validity

The first identifiable feature of the survey responses was the tendency to disapprove of the EI format. The number of respondents who only "somewhat agreed" that EI was an accurate assessment of their Portuguese speaking skills was more than double that of the OPIc (10 vs 4), and the number of respondents who "agreed" the same thing about the OPIc was much higher than the EI (19 vs 11). Additionally, 12 of the respondents stated that even though they agreed on some level that EI was an accurate assessment of their speaking skills, they preferred the OPIc. Similar patterns were found in responses to the following question that asked participants why they chose the OPIc over the EI; nine of them alluded in some way to the idea that the OPIc allowed them to create with the language:

- *I think that understanding and responding reports a better representation of fluency as a whole.*
- *I feel that imitation measures only a portion of skill, whereas open ended questions allow for a more accurate assessment of comprehension and conversation skills.*
- *I preferred the OPIc because it felt more like a genuine assessment of how much Portuguese speaking I was actually capable of doing on the fly, instead of parroting back sentences that I didn't necessarily understand.*
- *I actually conversed in the OPI, which is a better measure of fluency than just listening to sentences.*
- *I think coming up with your own things to say is more of a reflection of what you know, rather than imitating something.*

Elicited Imitation continues to be perceived as an exercise in memorization or short-term memory, and is therefore perceived in a negative light:

- *I think the test asking to repeat back is more a rapid memorization test than anything. One could argue that it tests your ability to hear and distinguish between words, but that is not the best way to test it in my opinion. If I took that same test in English, my native language, then I probably would have scored very similar to how I scored in Portuguese.*
- *It (OPIc) mimics real life speech rather than repeating phrases that may or may not be in your vocabulary. Communication is more important than memorization.*

However, in these responses several of the respondents noted that they recognized the need for comprehension in order to successfully complete the EI tasks. All of this data leads me to believe that much of the negative criticism surrounding EI testing is a result of a lack of information. There were even a number of responses in the survey that expressed a desire to understand the science behind EI:

- *The EI test was definitely interesting, and something that I hadn't done before. I would be interested to see how scoring goes on this test. I would like to see how proficient I was projected to be based on the EI.*
- *It's like the game of Simon. It's easy at first to just copy the sounds, but at point you have to actually memorize the words or ideas, not the sounds.*

My interpretation of these data is that learners seeking to assess their skills for any number of reasons would be more open to the EI process if they were given the opportunity to understand the science better. One participant made the following comment:

> *After having taken both tests, I asked the girls running the tests about them and they explained to me how they worked. The EI test is cool because in order to do well with the*

> *more complex sentences, one needs to have already passed that level of their comfort zone, showing fluency in speaking about like topics and with similar vocabulary. On the other hand, less experienced speakers will become overwhelmed with larger and more complex sentences and be unable to repeat them. That's why I feel the EI test is my preferred option.*

Upon my initial reading of this comment I was concerned that the participant's view of the tests had been changed by the comments of the administrator (an employee of Emmersion Learning), which was not supposed to happen before they took the survey and reflected on their experiences. However, this comment contains some revealing data that could be helpful in future implementation of EI tests. I do not know how this participant felt about the EI test prior to speaking with the administrator, but it appears that an understanding of the science behind the method helped this person to appreciate the potential results. Regarding the question of face validity, it is likely that the most effective tool in increased proliferation of EI testing methods will be information and education regarding the research and evidence supporting its use. This education will have to start with language educators.

Another recommendation for increasing the face validity of EI testing is the inclusion of some open-ended questions with the EI items. Obviously, automated grading of open-ended questions poses a significant challenge. Therefore, these may be ungraded questions that serve the purpose of putting the student at ease regarding the testing method or may be presented as a warm-up exercise. Currently, the final Emmersion Learning Portuguese Elicited Imitation test contains an open-ended question at the end that does not factor into the grade but allows the student to spend time creating with the language. There is a question of whether or not such an inclusion is ethical if students are led to believe that the response to the question will factor into

their results. EL is currently pursuing automatic speech recognition technology that could accurately incorporate the response to this question into the overall score.

The EI Experience

When asked to share any further thoughts regarding either of the tests there was an unexpected theme of overall discouragement expressed by those who were in the Novice levels. Some examples of this follow:

- *I believe it would be more productive to measure comprehension in a way that doesn't result in a universal lack of effort.*
- *I know it's just to see how we're doing with the language but I felt like I would never learn Portuguese after I took those.*
- *…it sounded very robotic and made me feel pretty sad about my Portuguese after 6 weeks in the MTC.*
- *The difficulty curve seemed randomly timed, and there was no point after the difficulty where it went to an easier point to restore confidence.*

Because I was one of the raters, I had the opportunity of listening to the examinees as they progressed through the EI test, which progressed from 5-7 syllable sentences up to 30-35 syllable sentences. This theme was very apparent in the tone of voice of the participants. Some of them were very clearly making nonsensical noise rather than even attempting to imitate some of the words they heard. At the beginning of the test, all of them were attempting to do well but very quickly became discouraged and were merely trying to get to the end.

This is an important advantage of the OPI or OPIc over an EI test. The interviewer in an OPI is able to adapt to the level of production of the examinee, and in the OPIc the questionnaire at the beginning will allow the test to be adjusted to appropriately fit the level of the examinee.

In order to avoid this type of examinee response in an EI examinee, it will be necessary that test creators employ adaptive automated scoring. Emmersion Learning is current employing Automated Speech Recognition (ASR) technology for grading their tests and working toward adaptive testing as well. Having listened to the responses of the examinees, I am confident that if they had been given the opportunity to respond to more Novice level items to build confidence, they would have done better than they did. Until adaptive technology is incorporated into EI testing, its use for assessing Novice level learners may be unwise as it may be a primarily negative experience for them, as shown by the responses to the survey. I will discuss a potential solution to this problem in the suggestions section at the end of this chapter.

Regarding the question of what level of stress was induced by the different tests, I was surprised to learn that nearly as many participants considered the EI test to be a stressful experience as the OPIc. I would have assumed the nature of the EI (listen and repeat) would have made it relatively stress free as the OPIc requires an examinee to consider the questions being asked and formulate responses. There were some comments in the written questions that confirmed my assumptions:

- *I could try and say what I heard which is easier than having to come up with an answer to something I didn't understand.*
- *The EI was good and less stressful.*

Further reading into the comments leads me to believe that some of the stress in the EI test was resulting from unfamiliarity with the format and perhaps some lack of clarity in the instructions. This information further confirms my earlier statement regarding the importance of a greater understanding of the EI process.

Limitations

After analyzing the results of this study, it is important to mention some of the limitations to the process. One that must be recognized is the original assumption of validity of the OPIc as a valid criterion by which to establish the EI test's validity. The criticisms of the OPI and OPIc have been discussed in previous chapters. However, their place in the current foreign language learning environment as the "gold standard" in language assessment makes the OPIc the most appropriate criterion for this study.

Secondly, due to a number of unforeseen circumstances, the survey was distributed several months after the tests had been administered and therefore resulted in a number of participants not being able to respond. Many of the novice level participants were trainees in the Missionary Training Center of The Church of Jesus Christ of Latter-Day Saints, and soon after they took the tests, they were sent to remote locations around the world where they would have been unable to respond to the survey due to any number of reasons. Additionally, the delayed distribution of the survey likely resulted in some lapses in participants' recollection of the EI and OPIc experiences.

Finally, the initial EI calibration experience does not reflect a true EI test for several reasons. First, a typical EI test will be taken on a personal computer in the privacy of one's home, office, or other preferred location. Examinees will likely be alone in a quiet environment. For the purposes of our study it was necessary to have the participants come to a computer lab at a scheduled time where they took the test together with other participants. Some of the participants complained of ambient noise and problems with the headsets they were given. Additionally, the calibration test that participants took was 60 items long as it was designed to identify the 30 best items that would be used to create the final test that would be

commercialized. A 60 item EI test is long and may have resulted in test fatigue for some and significant decrease in motivation to perform well, as well as being long and emotionally draining for anyone who was only able to successfully complete some of the first few items. This may have driven some of the negative comments received in the survey.

Conclusions and Recommendations for Further Research

Results of the aforementioned validation study suggest that the Portuguese EI test appears to be a valid oral proficiency assessment instrument. The criterion-related validity was established through statistical comparison of test results of the EI calibration test with those of the OPIc, a widely accepted assessment tool. There are a number of advantages to the EI test versus the OPI or OPIc. The most important advantage may be the affordability of the test. This is not only true for individuals interested in acquiring an assessment of their speaking ability, but also and perhaps especially to organizations that need to assess the skills of large numbers of employees, candidates, or students. In addition to being affordable, the EI testing method is also much more logistically simple. There is no need for scheduling, as EI examinees can take the test whenever and wherever they like as long as they have an internet connection and a computer with audio and a microphone. Results are immediate and come with comparisons to the other prevailing assessment scales such as ACTFL and CEFR.

It can also be concluded that there continues to be an issue of face validity associated with EI testing. Participants are suspicious of the format primarily because it appears on the outset to be a mere exercise in memorization and parroting. This study, in combination with many others, provides evidence to the contrary – that EI testing achieves many of the same results when compared specifically to ACTFL OPIc and inferentially through that comparison to the ACTFL OPI as well as the CEFR.

Extensive research has been completed on the different aspects of EI item creation, including the selection of lexical and morphological content, as well as different validity studies regarding specific EI tests. However, I was unable to locate research specifically related to the experience of test-takers and how different learners feel about the EI format. An area that should be addressed is the effect that personality plays in test method preferences as it relates specifically to EI but also to the OPI and OPIc. When the OPI was in its infancy, Young (1986) completed a study that showed significant negative correlation between anxiety and performance on the OPI. I would hypothesize that a different testing method such as the EI would provide interesting insight into oral proficiency assessments of those who may suffer from social anxiety in a scenario like an OPI. Although some individuals suffer from general text anxiety, regardless of the format, there are likely others who would perform differently in an environment where they are not being interviewed by a stranger, or even an avatar.

A second suggestion for further research is in relation to testing methods. Much of the information gleaned from survey responses leads me to believe that if participants had a better understanding of the research behind EI and the correlations between EI testing and interview style assessment, their overall attitudes towards the EI method would improve. A study of a control group with little knowledge of the science supporting EI methods and outcomes in comparison to a group that had been educated on the value and validity of EI might reveal an improvement in attitudes as well as outcomes.

Lastly, I recommend that EI test creators explore the possibility of creating a lower-level test that contains items that only reach the Intermediate Mid sublevel. I previously discussed the negative reactions of several Novice level speakers to the EI test format. Many of them had very discouraging experiences because of their inability to respond to the vast majority of the test

items. EI test creators could include a self-assessment questionnaire at the beginning of the exam which identified individuals that would be very unlikely to achieve a score any higher that Intermediate Mid. A second test containing only items appropriate to these levels could be finely tuned to more precisely identify the test-takers level. In the case of an individual achieving the highest possible outcome on this lower-level test the option of taking the upper-level exam could be provided at no extra charge, thereby providing a positive and more informative experience for the individual.

References

American Council on the Teaching of Foreign Languages. (2012). *Oral proficiency interview: Familiarization manual.* Alexandria, VA: Author.

American Council on the Teaching of Foreign Languages. (n.d.). *Oral proficiency assessments.* Alexandria, VA: Author. Retrieved from https://www.actfl.org/professional-development/assessments-the-actfl-testing-office/oral-proficiency-assessments-including-opi-opic

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, *16*(4), 449-465. https://doi.org/10.2307/3586464

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The Modern Language Journal,70*, 380-390. https://doi.org/10.2307/326817

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E. E. Tarone, S. M. Gass, & A. D. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 245-261). Milton Park, UK: Routledge.

Brindley, G. (1998). Describing language development? Rating scales and SLA. In L.F. Bachman & A.D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112-140). Cambridge, UK: Cambridge University Press.

Bowden, H. W. (2016). Assessing second-language oral proficiency for research: The Spanish elicited imitation task. *Studies in Second Language Acquisition, 38*, 647-675. https://doi.org/10.1017/S0272263115000443

Burdis, J. R. (2014). *Designing and evaluating a Russian elicited imitation yet to be used at the Missionary Training Center* [Unpublished doctoral dissertation]. Brigham Young University, Provo, UT.

Canale, M. (1983). 1983: From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2-27). London, UK: Longman.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*, 16-33. https://doi.org/10.1177/026553229501200102

Chalhoub-Deville, M., & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals*, *36*(4), 498-506. https://doi.org/10.1111/j.1944-9720.2003.tb02139.x

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87-185. https://doi.org/10.1017/S0140525X01003922

Cox, T. L., & Davies, R. S. (2012). Using automatic speech recognition technology with elicited oral response testing. *CALICO Journal, 29*, 601-618. https://doi.org/10.11139/cj.29.4.601-618

Dandonoli, P., & Henning, G. (1990). An investigation of the construct validity of the ACTFL proficiency guidelines and oral interview procedure. *Foreign Language Annals, 23*, 11-22. https://doi.org/10.1111/j.1944-9720.1990.tb00330.x

Davies, M., & Preto-Bay, A. M. R. (2007). *A frequency dictionary of Portuguese*. Milton Park, UK: Routledge.

Drackert, A. (2015). Elicited imitation. In R. Grotjahn & G. Sigott (Eds.), *Validating language proficiency assessments in second language acquisition research: Applying an argument-based approach* (pp. 49-65). Frankfurt, Germany: Peter Lang.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics, 27,* 464-491. https://doi.org/10.1093/applin/aml001

Fails, W., & Clegg, J. H. (2020). *Manual de fonética e fonologia da língua portuguesa*. Unpublished manuscript, Brigham Young University, Provo, UT.

Gaillard, S. (2014). *The Elicited Imitation Task as a method for French proficiency assessment in institutional and research settings* [Doctoral dissertation]. University of Illinois at Urbana-Champaign.

Gallimore, R., & Tharp, R. G. (1981). The interpretation of elicited sentence imitation in a standardized context. *Language Learning*, *31*(2), 369-392. https://doi.org/10.1111/j.1467-1770.1981.tb01390.x

Graham, C. R., McGhee, J., & Millard, B. (2010). The role of lexical choice in elicited imitation item difficulty. In M. T. Prior, Y. Watanabe, & S. Lee (Eds.), *Selected proceedings of the 2008 Second Language Research Forum* (pp. 57-72). Somerville, MA: Cascadilla Proceedings Project.

Herzog, M. (2003). Impact of the proficiency scale and the oral proficiency interview on the foreign language program at the Defense Language Institute Foreign Language Center. *Foreign Language Annals, 36*, 566-571. https://doi.org/10.1111/j.1944-9720.2003.tb02146.x

Higgs, T. V., & Clifford, R. (1982). The push toward communication. In T. V. Higgs (Ed.), *Curriculum, competence, and the foreign language teacher* (pp. 57-136). Lincolnwood, IL: National Textbook.

Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. *The Modern Language Journal*, *28*(2), 136-150. https://doi.org/10.2307/317331

Language Testing International. (n.d.). *ACTFL speaking assessment: The Oral Proficiency Interview-computer (OPIc)*. Retrieved February 20, 2020, from https://www.languagetesting.com/oral-proficiency-interview-by-computer-opic

Lantolf, J., & Frawley, W. (1985). Oral-proficiency testing: A critical analysis. *The Modern Language Journal, 69*, 337-345. https://doi.org/10.2307/328404

Leclercq, P., & Edmonds, A. (2014). How to assess L2 Proficiency? In P. Laclercq, A. Edmonds, & H. Hilton (Eds.), *An overview of proficiency assessment research*: *Measuring L2 Proficiency* (pp. 3-23). Bristol, UK: Multilingual Matters.

Liskin-Gasparro, J. E. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals,36*, 483-490. https://doi.org/10.1111/j.1944-9720.2003.tb02137.x

Lust, B., Chien, Y. C., & Flynn, S. (1987). What children know: Methods for the study of first language acquisition. In B. Lust (Ed.), *Studies in the acquisition of anaphora* (Volume II, pp. 271-356). Dordrecht, Netherlands: Springer.

Malone, M. E., & Montee, M. J. (2010). Oral proficiency assessment: Current approaches and applications for post-secondary foreign language programs. *Language and Linguistics Compass, 4*, 972-986. https://doi.org/10.1111/j.1749-818x.2010.00246.x

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 343-355. https://doi.org/10.1037/h0043158

Okura, E., & Lonsdale, D. (2012). Working memory's meager involvement in sentence repetition tests. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34, pp. 2132-2137). Retrieved from https://escholarship.org/uc/cognitivesciencesociety

Omaggio, A. C. (1983). Methodology in transition: The new focus on proficiency. *The Modern Language Journal*, *67*(4), 330-341. https://doi.org/10.2307/327063

Ommagio, A. (1986). *Teaching language in context*. Boston, MA: Heinle & Heinle.

Surface, E. A., Poncheri, R. M., & Bhavsar, K. S. (2008). *Two studies investigating the reliability and validity of the English ACTFL OPIc with Korean test takers* (ACTFL OPIc Validation Project Technical Report). White Plains, NY: American Council on the Teaching of Foreign Languages and Language Testing International. Retrieved from https://www.languagetesting.com/pub/media/wysiwyg/research/ACTFL-OPIc-English-Validation-2008.pdf

Swaffar, J. K., Arens, K., & Byrnes, H. (1991). *Reading for meaning: An integrated approach to language learning*. Englewood Cliffs, NJ: Prentice Hall.

Swender, E., & Vicars, R. (1999). *ACTFL oral proficiency interview tester training manual*. Yonkers, NY: American Council on the Teaching of Foreign Languages.

Thompson, G. L., Cox, T. L., & Knapp, N. (2016). Comparing the OPI and the OPIc: The effect of test method on oral proficiency scores and student preference. *Foreign Language Annals, 49*, 75-92. https://doi.org/10.1111/flan.12178

Tigchelaar M. (2019). Exploring the relationship between self-assessments and OPIc ratings of oral proficiency in French. In P. Winke & S. Gass (Eds.), *Foreign language proficiency in higher education* (Educational Linguistics, Vol. 37, pp. 153-173). Cham, Switzerland: Springer International. https://doi.org/10.1007/978-3-030-01006-5_9

Tracy-Ventura, N., McManus, K., Norris, J. M., & Ortega, L. (2014). 'Repeat as much as you can': Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, H. Hilton, & A. Edmonds (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 143-166). Bristol, UK: Multilingual Matters. https://doi.org/10.21832/9781783092291-011

TrueNorth. (2019). *Technical report: Portuguese TrueNorth Test (TNT) Form A*. Lehi, UT: Author.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*(3), 325-344. https://doi.org/10.1177/0265532211424478

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*(1), 54-73. https://doi.org/10.1111/1473-4192.00024

Wu, S., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals, 46*, 680-704. https://doi/org/10.1111/flan.12063

Young, D. J. (1986). The relationship between anxiety and foreign language oral proficiency ratings. *Foreign Language Annals*, *19*(5), 439-445. https://doi.org/10.1111/j.1944-9720.1986.tb01032.x

APPENDIX A

Participant Test Experience and Reaction Survey

1. The elicited imitation test was an accurate assessment of my Portuguese speaking ability.

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree

2. The Computerized Oral Proficiency Interview was an accurate assessment of my Portuguese speaking ability.

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree

3. Given the two options, which of the two tests would you prefer to take as an assessment of your Portuguese speaking skills.

- Oral Proficiency Interview by Computer
- Elicited Imitation
- No preference

4. Would you please explain in more detail the reason for your selection in question 2 above:

5. Please share any other thoughts on either the EI test or OPIc.

6. The instructions for taking the test were easy to understand.

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree

7. Taking the OPIc was a stressful experience.

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree

8. Taking the EI Test was a stressful experience.

Strongly Disagree	Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree

9. If EI were to be used in the future to assess students' Portuguese speaking levels, what suggestions, if any, would you have for improving the test?

10. Did you experience any technical difficulties on either of the tests?

- Yes
- No

11. If you answered "Yes" above, please give details as to your technical difficulty: