2017

Spanish listening comprehension test development project: Considering learner L2 listening comprehension in the course placement equation

Jennifer Elise Musgrove

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Bilingual, Multilingual, and Multicultural Education Commons, English Language and Literature Commons, and the Linguistics Commons

Recommended Citation

Musgrove, Jennifer Elise, "Spanish listening comprehension test development project: Considering learner L2 listening comprehension in the course placement equation" (2017). Graduate Theses and Dissertations. 16091.

https://lib.dr.iastate.edu/etd/16091

This Thesis is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Spanish listening comprehension test development project: Considering learner L2 listening comprehension in the course placement equation

by

Jennifer Elise Musgrove

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

Major: Teaching English as a Second Language/Applied Linguistics (Teaching English to L1 Spanish Learners)

Program of Study Committee:
Gary Ockey, Major Professor
Chad Gasta
Evgeny Chukharev-Khudilaynen

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this thesis. The Graduate College will ensure this thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2017

Copyright © Jennifer Elise Musgrove, 2017. All rights reserved.
DEDICATION

Para mi hombre, Wesley, mi mejor amigo, cuyo consejo para animarme “How do you eat an elephant?...One bite at a time!” evidentemente me ha servido muy bien. Y, para mis hijos Forrest Emerson y Eli Thomas. Como vuestra mamá, disfruté mucho de compartir la experiencia única de ser una estudiante al lado vuestro.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>v</td>
</tr>
<tr>
<td>NOMENCLATURE</td>
<td>vi</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>vii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>viii</td>
</tr>
<tr>
<td>CHAPTER 1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Iowa State University Spanish language placement:</td>
<td></td>
</tr>
<tr>
<td>The WebCAPE placement exam</td>
<td>1</td>
</tr>
<tr>
<td>WebCAPE placement score concerns</td>
<td>2</td>
</tr>
<tr>
<td>Rationale for the development of a L2 listening comprehension test</td>
<td>3</td>
</tr>
<tr>
<td>Research questions</td>
<td>4</td>
</tr>
<tr>
<td>CHAPTER 2 WEBCAPE TEST REVIEW AND LITERATURE REVIEW</td>
<td>6</td>
</tr>
<tr>
<td>Creating a more authentic placement exam</td>
<td>10</td>
</tr>
<tr>
<td>Approach</td>
<td>11</td>
</tr>
<tr>
<td>CHAPTER 3 METHODS</td>
<td>13</td>
</tr>
<tr>
<td>Participants</td>
<td>13</td>
</tr>
<tr>
<td>Procedures</td>
<td>14</td>
</tr>
<tr>
<td>Test format and delivery options</td>
<td>15</td>
</tr>
<tr>
<td>Test pilot phase one:</td>
<td>17</td>
</tr>
<tr>
<td>Preliminary test review and instructor feedback</td>
<td></td>
</tr>
<tr>
<td>Test pilot phase two:</td>
<td>18</td>
</tr>
<tr>
<td>Test-taker participants</td>
<td></td>
</tr>
<tr>
<td>CHAPTER 4 RESULTS AND DISCUSSION</td>
<td>20</td>
</tr>
<tr>
<td>CHAPTER 5 SUMMARY AND CONCLUSIONS</td>
<td>33</td>
</tr>
<tr>
<td>Closing acknowledgements</td>
<td>35</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>36</td>
</tr>
<tr>
<td>APPENDIX A: FEEDBACK CONSENT FORM</td>
<td>38</td>
</tr>
<tr>
<td>APPENDIX B: HUMAN SUBJECT CONSENT FORM</td>
<td>40</td>
</tr>
<tr>
<td>APPENDIX C: IRB APPROVAL</td>
<td>43</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>WebCAPE test item</td>
<td>7</td>
</tr>
<tr>
<td>Figure 2</td>
<td>WebCAPE test item</td>
<td>8</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Example test item</td>
<td>15</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Screenshot of L2 Listening Comprehension Exam</td>
<td>16</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Histogram of test-taker scores</td>
<td>21</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Screenshot of test item 14</td>
<td>29</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Statistical calculations for L2 Spanish Listening Comprehension Exam</td>
<td>22</td>
</tr>
<tr>
<td>Table 2</td>
<td>L2 Spanish Listening Comprehension Exam discrimination analysis values</td>
<td>23</td>
</tr>
<tr>
<td>Table 3</td>
<td>Item facility scores for test items 1-8 within L2 Spanish Listening Comprehension Exam</td>
<td>24</td>
</tr>
<tr>
<td>Table 4</td>
<td>Item facility scores for test items 9-15 within L2 Spanish Listening Comprehension Exam</td>
<td>25</td>
</tr>
<tr>
<td>Table 5</td>
<td>Point-biserial scores for test items 1-8 within L2 Spanish Listening Comprehension Exam</td>
<td>26</td>
</tr>
<tr>
<td>Table 6</td>
<td>Point-biserial scores for test items 9-15 within L2 Spanish Listening Comprehension Exam</td>
<td>26</td>
</tr>
<tr>
<td>Table 7</td>
<td>Response point-biserial scores for test items within L2 Spanish Listening Comprehension Exam</td>
<td>28</td>
</tr>
</tbody>
</table>
NOMENCLATURE

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>First Language</td>
</tr>
<tr>
<td>L2</td>
<td>Second Language</td>
</tr>
<tr>
<td>SLA</td>
<td>Second Language Acquisition</td>
</tr>
<tr>
<td>NRT</td>
<td>Norm-referenced Test</td>
</tr>
<tr>
<td>CRT</td>
<td>Criterion-referenced Test</td>
</tr>
<tr>
<td>SEM</td>
<td>Standard Error of Measurement</td>
</tr>
<tr>
<td>TLU</td>
<td>Target Language Use</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

First and foremost, I would like to thank my committee chair, Dr. Gary Ockey, and my committee members, Dr. Chad Gasta, and Dr. Evgeny Chukharev-Khudilaynen, for their guidance and support throughout not only the course of this research project but my entire graduate experience. Dr. Evgeny Chukharev-Khudilaynen motivates his students to stretch beyond their perceived potential and his introduction to linguistics course served to confirm that I am a linguist at heart who can appreciate the wit behind his wishing students to “have an average day”. Dr. Gary Ockey’s Second Language Assessment course served to spark an interest in an SLA area of research I had no prior knowledge or experience with. Through Dr. Ockey’s expert instruction delivered with a sense of humor and willingness to go above and beyond for his students I was able to grasp the complex discipline of second language assessments and move confidently forward with this research. Furthermore, this research would not have been possible without Dr. Chad Gasta, World Languages and Cultures Dept. Chair, whose continual endeavor to provide the highest quality language learning experience for WLC students is an inspiration and has made a tremendous impact on my personal second language learning journey.

I would also like to extend a “muchísimas gracias” to ISU Department of World Languages and Cultures and the following WLC colleagues for their enthusiastic participation, time, support, professional feedback, and encouragement; Elisa Rizo, Julia Domínguez, Marta Lence, Carly Johansen, Charles Nagle, Erik Ladner, Celeste Gonzalez-Chaves, Thomas Waldemer, Lucía Suárez, Flor Slowing, and Megan Myers. Also, I would like to thank our motivated and fantastic WLC Spanish students for their part as test-takers. Lastly, I would like to acknowledge Forrest Musgrove for sharing his technical knowledge and assistance with the video coding in order to provide high-quality listening assessment material for this project.
ABSTRACT

Second language (L2) placement exams utilized towards course placement decisions at higher-level educational institutions are considered “high-stakes” relative to the impact that the assessment outcome has on the individual student level as well consequences at the institutional level. Therefore, providing reliable second language assessments that effectively determine L2 learner language proficiency can lower course placement error issues. I argue that assessing a learner’s L2 listening comprehension is an integral part of the learner’s L2 experience within the Target Language Use (TLU) and that it could improve course placement accuracy. Furthermore, I argue for a listening construct that includes non-verbal communication and facial cues that videos provide in order to create an authentic test experience. Utilizing this construct, I then created an L2 Spanish listening comprehension exam utilizing authentic video material and created test items based on Brown (2005) multiple-choice guidelines. After incorporating feedback from the first phase of the test pilot, as the second phase I then administered the exam as a diagnostic the first two weeks of Fall 2017 semester to over four hundred currently enrolled WLC Spanish language students comprised of beginning through advanced language learners. Utilizing Classical Test Theory to guide my statistical analysis of test scores, results reveal a normal distribution of scores that parallels the representative distribution of test-taker levels. Moreover, the Cronbach’s alpha reliability estimate of 0.73 reflects relatively good reliability. Item discrimination as well as distractor analysis reveal acceptable test items. Overall, this study suggests that it is possible to create a reliable multiple-choice video-delivered listening test for Spanish users that could conceivably be used to accurately place students into appropriate classes.
CHAPTER 1. INTRODUCTION

In seeking to provide second language (L2) learners with the highest quality language learning experience at the university level, the preliminary step is to consider which course best meets the learner’s needs in connection with their present L2 learner level. In order to properly assess students’ L2 language skills, higher-level education institutions can develop or purchase access to placement exams for the target language and utilize the generated placement scores towards course placement recommendations. However, much is at stake for the learner when assessing an individual’s second language competence for university level course work. What are the consequences for the learner if placed beyond their language level? Conversely, what are the consequences if placed beneath their language level? Therefore, developing and administering an L2 assessment that effectively identifies L2 learner level with limited placement error can be considered a crucial component of a successful language learner experience.

Iowa State University Spanish language placement: The WebCAPE placement exam

The WebCAPE Spanish placement exam, created in 2001, was created by Brigham Young University’s (BYU) Humanities Technology and Research Support Center and is managed by Perpetual Technology Group. WebCAPE (Computer Adaptive Placement Exam) is primarily designed for university level students with previous L2 experience as the target population. Currently, Iowa State University World Languages and Cultures Department (WLC) provides the WebCAPE Spanish placement exam to all current or potential students as a starting point in the process of course registration. Students log on via the language department’s main web page in order to take the exam and scores are then automatically delivered to both the
student and to WLC academic advisors. Instructions on the WLC website encourage students to make an appointment with an advisor after taking the exam in order to discuss their current Spanish language proficiency and recommended course placement. Using score classifications provided by WebCAPE, test-takers are assigned a raw score of 0-300 as elementary/beginning level, a score of 300-400 as low-intermediate level, and a score of 400+ as intermediate level Spanish. However, WebCAPE scores are used in addition to any previous experience the learner may have with the language including past high school coursework, and/or living abroad in a Spanish speaking country. For example, if a student receives a raw score close to 300 but also has previous high school coursework or study abroad experience, they may be placed in a low-intermediate rather than beginning level course.

**WebCAPE placement score concerns**

While WebCAPE has been generally effective for its purpose, some concerns have been raised. The WebCAPE website cites a BYU study that found 143 out of 179 (79.9%) of the students that participated in the placement exam were placed appropriately. Out of the approximately 20% of students that were improperly placed, the study found that the majority should have been placed one course higher, reflecting mostly conservative placement errors (“Details;Validation”, 2016). Taking this into consideration, throughout the course of Spring 2017 semester I formally interviewed WLC Spanish instructors with experience teaching beginning, intermediate, and advanced levels as well as WLC academic advisors. I asked them to share their experiences related to Spanish language course placement for language learners newly entering the program or existing students enrolling in courses. The consensus among instructors and academic advisors is that it appears there is a general trend of WebCAPE scores indicating language ability beyond what the language learner is actually capable of, especially
with regards to listening comprehension and speaking tasks within the formal classroom environment. Additionally, some Spanish L2 instructors expressed ongoing concerns of “losing students” who were improperly placed. They explained that once it becomes clear to a student and/or their instructor that the student needs to transfer to the appropriate course level, all course sections are typically full or semester coursework has progressed far enough that it is unadvisable to transfer to a different course. Unfortunately, the student then must decide if they are capable of continuing or instead need to drop the course and wait to enroll in the appropriate level the following semester. In addition to instructor and advisor feedback, I then conducted a test evaluation of the WebCAPE exam in order to determine the test’s usefulness as a Spanish language placement exam in connection with current SLA literature related to L2 listening comprehension and assessment.

**Rationale for the development of a L2 listening comprehension test**

The purpose of this study was to create a reliable Spanish language listening comprehension test that assesses learner L2 language comprehension comprised of beginning, intermediate, and advanced level Spanish language students that could be used to increase the placement reliability in a Spanish program. A preliminary step in the test development process was to gain instructor and academic advisor feedback as well as conduct a test review of the WebCAPE exam, the Spanish placement exam currently utilized within Iowa State University’s Dept. of World Languages and Cultures (WLC). Preliminary WebCAPE test review results, a second language acquisition (SLA) literature review focused on listening comprehension, as well as instructor and advisor feedback guided the process of creating a listening comprehension assessment in order to address the following research questions based upon Classical test theory.
Research questions

The overarching aim of this study was to determine the extent to which an L2 video-delivered Spanish listening assessment can be considered a reliable assessment of L2 listening comprehension. To fulfill this purpose, I focused on two specific research questions related to the reliability of this test:

Research Question 1: How well do test items measure test-takers’ L2 listening comprehension skills when based upon video recorded authentic language discourse that incorporates facial cues and body language into the test construct?

Research Question 2: What is the reliability of the test as measured by Cronbach’s alpha?

In order to address these research questions, I used descriptive statistics, classical test theory item level analysis, and Cronbach’s alpha to determine how well the test distinguishes test-takers, i.e. how well the test separates out test-takers across an L2 listening comprehension ability range when facial cues and body language are incorporated into the test construct. Additionally, the alpha reliability estimate was utilized to calculate the Standard Error of Measurement (SEM) which demonstrates the reliability of individual test-taker scores.

In this chapter, I have indicated the problem and described how I will address it in this thesis. In chapter 2, I will discuss results from a WebCAPE test review I conducted based on Bachman & Palmer’s (1996) Test Usefulness Framework in connection with a review of SLA literature related to L2 listening comprehension and assessment. In chapter 3, I will discuss test development procedures, test participants, and test delivery as a diagnostic.
In chapter 4, I will present and discuss the results of the study. Lastly, in chapter 5, I will summarize the study results and provide suggestions pertinent to future implementation of the test.
CHAPTER 2. WEBCAPE TEST REVIEW AND LITERATURE REVIEW

Following the test usefulness framework introduced by Bachman & Palmer (1996) and modified by Stoynoff & Chapelle (2005), I conducted an evaluation of the WebCAPE placement exam based on test authenticity and possible impact, two important aspects of validity connected to L2 test development.

Authenticity refers to the relationship between the test’s tasks and the Target Language Use (TLU) within the domain. A test is described as “relatively more/less authentic” according to how successfully the tasks correspond with the TLU domain. In this case, the TLU is the particular Spanish language course that the student would enroll in after taking the placement exam. For example, Spanish 303A (Spanish Conversation and Composition: Through Culture) and Spanish 303B (Spanish Conversation and Composition: For Professionals) are advanced level courses currently offered at Iowa State University within the Department of World Languages and Cultures. Instruction is entirely in Spanish, which requires at least an entry level advanced-intermediate level of reading, writing, speaking, and listening skills upon starting the course. In the classroom, students are required to communicate with the instructor and one another exclusively in the target language. In addition to assigned readings, written coursework, and group projects, assignments include watching topic related videos with follow-up small group and/or entire class discussions in the target language. Therefore, assessment tasks that corresponded with the previously described TLU would create a relatively more authentic placement exam.

In comparison, WebCAPE claims that the exam’s construct used to measure the test-taker’s current L2 proficiency and fluency skills to be utilized for course placement is based on
the American Council for the Teaching of Foreign Languages (ACTFL) proficiency guidelines. ACTFL guidelines separate language learners into four main placement categories: novice (low, mid, high), intermediate (low, mid, high), advanced (low, mid, high), and superior. Within each of these proficiency levels, ACTFL provides a list of what a learner “can” and “cannot” do within the L2.

The WebCAPE exam administers multiple-choice tasks, as shown in Figure 1 and Figure 2, that aim to test L2 grammar knowledge, vocabulary knowledge, and reading comprehension. Both Figure 1 and Figure 2 demonstrate test format, including test instructions provided in English and included with each displayed test item.

![Mark the letter of the response that most closely corresponds to the Spanish.](image)

Figure 1. WebCAPE test item. This figure illustrates an example of multiple-choice test items on WebCAPE exam.

Figure 1 illustrates a question prompt that focuses on testing the L2 Spanish learner’s knowledge related to Spanish idioms, which can widely vary in meaning based upon region and dialect. Additionally, this item reflects a portion of WebCAPE test items that utilize “A and B”, “All of the above”, “None of the above”, etc. as distractors (incorrect responses). Brown (2005)
multiple-choice guidelines state that these types of distractors should be avoided due to risk of increasing a test-taker chance of answering the item correctly when guessing (49-50). Figure 2 illustrates an example of cloze test items, also referred to as “fill-in-the-blank” test items, included in the WebCAPE exam.

![Multiple-choice question](image)

**Mark the letter of the best response or completion.**

Peter looks fat.
Pedro ____ gordo.

A  mira  
B  está  
C  se es  
D  se parece

Figure 2. WebCAPE test item. This figure illustrates an example of multiple-choice test items on WebCAPE exam.

Figure 2 displays a question prompt that focuses on testing the L2 Spanish learner’s ability to translate a simple sentence from English to Spanish via selecting the appropriate verb in Spanish that corresponds with “looks” in English within the context of one’s “appearance”. Therefore, although the exam is comprised of grammar, vocabulary, and reading comprehension questions, the WebCAPE exam does not provide tasks that measure L2 oral ability or listening comprehension. Therefore, the L2 proficiency scores generated from the exam are limited to the learner’s knowledge of grammar, vocabulary, and reading comprehension.

Wagner (2014) points out that although approximately 50% or more of a person’s time is spent listening during a communicative event, due to L2 listening assessment challenges, “the assessment of listening has historically been somewhat neglected and even overlooked in the
language assessment literature” (1). Additionally, second language acquisition (SLA) research points to “a considerable overlap between listening and reading ability” (Buck, 2001, p. 31). Thus, due to current L2 listening assessment challenges as well as the acknowledged overlap between listening and reading ability, why should we bother testing a learner’s L2 listening comprehension? Buck (2001) affirms that spoken language is comprised of unique characteristics not present in written language; “firstly, speech is encoded in the form of sound; secondly it is linear and takes place in real time, with no chance of review; and thirdly, it is linguistically different from written language” (4). While written language often undergoes the process of revision and the reader is presented with a “final draft” to encode, the spoken language can be thought of as a messier “rough draft” that requires additional L2 encoding skills. For example, spoken language consists of, “short phrases or clauses, called idea units, strung together in a rather loose way, often connected more by the coherence of the ideas than by any formal grammatical relationship” (Buck, 2001, p.9). Surrounding these idea units, spoken discourse utilizes vocabulary repairs, filled and unfilled pauses, false starts, and repeats. Spoken language also undergoes phonological modifications depending on the rate of speech or dialect of the speaker. Therefore, although spoken and written discourse may share some overlapping qualities, the additional characteristics unique to spoken language present the L2 learner with comprehension challenges not present in written form. The assessment of L2 listening comprehension reveals at what level the learner is capable of encoding these characteristics unique to spoken language. Moreover, listening comprehension is an integral part of the TLU within the L2 classroom. Therefore, the WebCAPE placement exam can be considered relatively less authentic due to a lack of correspondence with the TLU through assessment tasks limited to
grammar, vocabulary, and reading comprehension while omitting tasks that assess the learner’s level of L2 listening comprehension.

Impact refers to the potential positive or negative effect a test could have at both the micro level, such as the test-takers, as well as the macro level, such as an institution or an entire society. With regards to the potential negative impact of the WebCAPE exam for WLC students, there continues to be an issue with the portion of test-takers that receive a raw score that improperly places them in a course beyond their L2 level. Consequently, the negative washback is that students may feel overwhelmed upon starting the course and, as discussed earlier, eventually may drop the course, forfeiting their L2 learning experience. At the macro level, WLC utilizes past and present student enrollment in order to determine which courses to offer as well as how many sections of each course should be provided per semester. Information from interviews with instructors and advisors suggest a negative test washback towards ensuring an adequate number of course sections based on the representative L2 levels of WLC Spanish language learners. Moreover, this may negatively impact the number of ISU graduates who are able to complete a Spanish minor or double major within their undergraduate time frame.

**Creating a more authentic placement exam**

The preliminary step towards providing a more authentic WLC placement exam is to begin with a test development project. The scope of this particular project was to create an L2 Spanish listening comprehension test based on current SLA research related to L2 listening comprehension and assessment. Quantitative data gathered from test-takers comprised of beginning, intermediate, and advanced L2 Spanish language learners currently enrolled in a Spanish course through Iowa State University, as well as feedback from WLC Spanish instructors and advisors will provide direction for future Spanish placement exam development.
**Approach**

The first step in test development is to define the construct of the skill that will be assessed. For listening comprehension, test developers have four tasks to choose from depending upon the listening construct; audio-only, that focuses on aural ability alone; audio with visual stimulus including content and/or context based visuals, that incorporates visual stimuli via still images or photos; and video, that includes non-verbal communication such as facial cues and body gestures. Current SLA research related to assessing listening comprehension (Sueyoshi & Hardison, 2005; Ockey, 2007; Ockey, 2009; Wagner, 2010; Powers, 2011; Suvorov, 2013; Wagner, 2014; Lee, Lee, Liao, & Wang, 2015) reveals that there continues to be a lack of research related to the effects of visuals on listening tests. Due to this lack of empirical research the incorporation of visuals into listening comprehension assessments continues to remain limited and researchers remain divided as to how to define the construct for listening assessment. While some second language assessment researchers argue for a listening comprehension construct that includes the ability to process both visual and verbal information, others contend that the construct should include aural ability only as visual information is extraneous to listening comprehension. However, keeping listening test authenticity in mind, I argue that the listening comprehension construct for this project will include non-verbal communication and facial cues that video provides in order to align as closely as possible to the TLU, the classroom experience. Additionally, test-takers can experience uneasiness and tension related to L2 listening comprehension test anxiety that can negatively affect listening comprehension scores. Compared
with audio-only test tasks, visual stimulus has been found to reduce test-takers’ anxiety (Powers, 2011; Lee, 2015). Therefore, a video format for this project additionally takes into consideration the importance of the test-taker’s interaction with the test and limiting L2 listening assessment anxiety as much as possible.
CHAPTER 3. METHODS

I will now move on to an explanation of the process of selecting speakers for developing video material as well as a description of test-taker participants. Additionally, I will explain the procedures involved in the test development process including the development of video material, test format decisions, and test delivery options. Lastly, I will explain the preliminary test review within the first phase of the development process and test delivery as a diagnostic within the second phase of the development process.

Participants

Speakers

To begin the process of creating the listening comprehension assessment and in order to create video material authentic to the TLU language instruction domain, I recruited four participants comprised of four separate speech varieties within the Spanish language. Three participants were native Spanish-speaking instructors within WLC representing Argentina, Spain, and Mexico. The fourth participant was an advanced-level second language speaker of Spanish with experience living abroad in Spain. Therefore, speaker participants were comprised of both L1 and L2 Spanish speakers with varying speech varieties of Spanish in order to account for speech variation within the TLU (Buck, 2001; Ockey & French, 2016; Ockey, Papageorgiou, & French, 2016).

Test-taker participants

Test-taker participants were comprised of over 400 Spanish language learners enrolled in beginning through advanced Spanish courses offered through the Department of World Languages and Cultures, Iowa State University. Test-taker participants and their role during the
second phase of test development are discussed in greater detail within the “Test pilot phase two: Test-taker participants” section.

**Procedures**

**Video material development and test format**

In order to maintain authenticity of the video recordings, during each filming event participants were paired up and each group received the direction that they were to converse with each other about a topic that came spontaneously to mind with no script provided. Additionally, they understood that there would be only one take of the video as subsequent takes could lead to a polished, less authentic conversation. Additionally, filming took place in an area with limited background noise interference in order to ensure good sound quality via an external microphone placed on a coffee table between the speakers. During the editing process, the audio portion from the video camera was removed and the high-quality audio portion from the external microphone was spliced into the video and then seamlessly aligned via iMovie. Other than the audio splicing, the video material did not receive any other editing such as editing out filled or unfilled pauses that naturally occurred during the conversation. Next, in order to meet Institutional Review Board (IRB) regulations for test delivery and storage of test-taker scores, a course was created on Blackboard specifically for the delivery of the Spanish L2 listening comprehension exam (see Appendix C). In order to embed the video material into Blackboard for a compatible and seamless video-audio delivery, the videos were first uploaded to YouTube and then embedded into separate text boxes within Blackboard. Additionally, in order to keep the embedded videos
from displaying the “suggested videos” tab upon completion, it was necessary to add an HTML5 flag to remove them.

**Test format and delivery options**

Once the video portion was embedded into the test, multiple-choice listening comprehension test items were created based on content provided by the video material that measured both implicit and explicit L2 comprehension focus (Buck, 2001, p.18-24) based on Carr (2011) and Brown (2005) multiple-choice guidelines. Figure 3 is an example test item from the L2 listening comprehension exam that assesses explicit comprehension which includes plausible (although incorrect) distractors connected to the conversation in the second video.

![Figure 3. Example test item. This figure illustrates a multiple-choice item from the L2 Spanish Listening Comprehension Exam.](image)

Figure 3 illustrates Carr (2011) and Brown (2005) multiple-choice guidelines regarding avoiding distractors such as “a and b” or “none of the above”, providing plausible distractors, and ensuring that all options are carefully worded to avoid ambiguity (Brown, 2005, pg. 47).

Additionally, Figure 3 illustrates an example test item testing learner ability to process details discussed within the listening event. In this case, the speakers discuss a yearly school dance event in Argentina and how it is comparable to a prom due to electing a “king” and “queen”. The distractors are connected to details previously discussed by the speakers related to school
celebrations for spring in Argentina. Seeking to closely align with the TLU, each test item was developed to test inferencing ability for L2 listening (implicit L2 listening skills) or learner L2 ability to process details of a listening event (explicit L2 listening skills). In the interest of future test development, a full disclosure of test items will not be included.

Each test item was assigned the value of one point with a total of 15 test items or 15 points. In order to separate test-taker data by language level, an additional question was added as the first item on the exam that asked test-takers to identify which Spanish course they were currently enrolled in, which was not assigned a point value. As Figure 4 illustrates, after adding the test items into the Blackboard exam, detailed instructions for the test were created in English rather than in Spanish in order to account for the anticipated test-taker L1 to be predominately English.

Figure 4. Screenshot of L2 Spanish Listening Comprehension Exam. This figure displays example test video format and test taking instructions.
Test instructions were intentionally written in English rather than in the target language, Spanish, as the construct for this exam does not include measuring the ability to process test instructions in the target language. Additionally, Figure 4 illustrates that test instructions included the direction to listen exactly twice through each video while viewing and answering test items in order to align with the current TLU classroom language instruction domain which includes the same guidelines for in-class L2 listening activities and assessments. Lastly, test delivery options were set to include a password to ensure that test-takers could not access the test until their assigned test date in the lab. As a follow-up, the password was changed after each session to discourage sharing the password with subsequent test-takers from other course sections. Additionally, the test option was selected to show total score only rather than to show correct responses for incorrectly answered items to discourage the sharing of correct answers with other test-takers.

Test pilot phase one: Initial test review and instructor feedback

Before administering the test to designated test-taker participants (see subsequent section “Test pilot phase two; Test-taker participants”), I shared a draft of the Spanish L2 listening comprehension exam created in Blackboard with WLC Spanish instructors, academic advisors, and colleagues comprised of beginning L2 Spanish level speakers through native Spanish speakers (see Appendix A). A total of 8 participants took the exam and then followed up by sharing feedback related to test instructions, test format, audio and video quality, and level of engagement of video material. Feedback from participants was utilized towards the final draft of the assessment delivered to test-taker participants for phase two.
Test pilot phase two: Test-taker participants

Moving on to the second phase of the test development process, Spanish courses offered through the Department of World Languages and Cultures at Iowa State University, from beginning through advanced Spanish (Spanish 101- Spanish 303A/B) incorporated the L2 Spanish listening comprehension assessment into course syllabi which was delivered as a diagnostic exam within the first two weeks of the Fall 2017 semester. Approximately 430 students total currently enrolled at the start of the semester in a WLC Spanish course took the exam during their designated class period in the Cargill Lab located in Pearson Hall (see Appendix B). Cargill Lab offered 30 stations of newly installed computers and fully-adjustable headsets to ensure the highest quality of audio and visual delivery of the exam. After explaining the purpose of the exam and answering all questions related to test instructions and test delivery, students were instructed to log into Blackboard and take the exam at their individual pace as I remained present in the lab to monitor test-takers, answer questions, and problem solve any technical issues that could arise. After test-takers completed the exam and reviewed their scores, I conducted a follow-up informational session related to listening comprehension and how to develop learner L2 listening comprehension skills based on guidelines for L2 learners provided by Fenner (2011).

As part of the test delivery process, since the Spanish L2 listening comprehension exam was delivered as a diagnostic exam at the start of the semester via Blackboard, students were enrolled by the registrar’s office into the specific “Spanish L2 Listening Comprehension Test Development Project” course in Blackboard by utilizing course lists of enrolled students for Spanish 101 through Spanish 303A/B. Due to a registration entry error, one section of Spanish 303B was not enrolled and therefore was not able to access the exam on their designated test day.
In order to benefit from the diagnostic exam and informational session, this particular section was able at a later date to access the test and L2 listening comprehension informational material. However, only 7 students total opted to take the test. These 7 scores were dropped from data set in order to maintain data integrity. Following the data gathering phase, I then conducted a descriptive statistical analysis of the data to establish test reliability in order to address Research Question 1 and Research Question 2.
CHAPTER 4. RESULTS AND DISCUSSION

Before discussing results from the descriptive statistical analysis, I will outline the difference between Norm-referenced tests (NRT) and Criterion-referenced tests (CRT) in order to provide background information related to the statistical analysis for this particular Spanish L2 listening comprehension exam. Norm-referenced tests (NRT) are tests whose scores are used to compare test-taker scores among themselves. These scores are commonly represented as percentiles. NRT scores will align on a graph in the shape of a curve that reflects an “average”, “above average” or “below average” percentile range. For example, if a student scored a percentile rank of 80, this is interpreted that the student scored higher than 80 percent of the students within the average group, but that 20 percent of students scored as well or better.

Criterion-referenced tests (CRT) differ from NRT’s in that scores are not generated through comparison among test-takers, but rather how well the test-taker met the standards or established objectives (Carr, 2011). This Spanish L2 listening comprehension exam utilizes the NRT paradigm to compare learner performance within the test-taker population in order to determine the extent to which this exam can be considered a reliable assessment of L2 listening comprehension. Thus, the forthcoming statistical analyses pertain to the NRT paradigm.

In assessments, a histogram is a visual representation that can be utilized to provide a descriptive statistical analysis of test-taker scores. Figure 5 displays the frequency of test-taker
scores and can be utilized to visually address Research Question 1: How well do test items measure test-takers’ L2 listening comprehension skills when based upon video recorded authentic language discourse that incorporates facial cues and body language into the test construct?

![Frequency of test taker scores](image)

Figure 5. Histogram of test-taker scores. This figure visually represents test takers’ score distribution of L2 Spanish Listening Comprehension Exam.

As the histogram visually displays, there is a spread of scores ranging between 1 to 15 points (out of 15 possible points). The total number of test-takers were fairly evenly distributed between levels, comprised of beginning, intermediate, and advanced learners. However, within the intermediate level test-takers were comprised of both intermediate (Spanish 201) and advanced-intermediate (Spanish 202, Spanish 297). Therefore, the distribution of test taker scores parallels the representative distribution of test-taker levels.

An additional statistical analysis related to the test items and test-taker responses provides the opportunity to further address Research Question 1. Table 1 displays total sample size as n equaling 414 test-takers. The minimum (Min.) score was a total of 1 point and the maximum score was a total of 15 points. The mean, median, and mode were similar: 9.13, 9.00,
and 11.00, respectively. The standard deviation (SD) was 3.01 with a skew of -0.36 and kurtosis of -0.45.

Table 1

<table>
<thead>
<tr>
<th>n</th>
<th>Min.</th>
<th>Max.</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>SD(Pop)</th>
<th>Skew</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>414</td>
<td>1</td>
<td>15</td>
<td>9.13</td>
<td>9.00</td>
<td>11.00</td>
<td>3.01</td>
<td>-0.36</td>
<td>-0.45</td>
</tr>
</tbody>
</table>

When the mean, median, and mode values of the exam scores are similar, it can be considered “normal” data, or a normal distribution (Ockey, 2016). With respect to this exam, since the mean, median, and mode are fairly similar it can be described as reflecting a relatively normal distribution of scores. The standard deviation of the population refers to the average difference between individual scores and the mean when we wish to focus on the entire group that took the test rather than generalizing to a larger population (Carr, 2011). A standard deviation of 3.0 can be considered relatively large in comparison to the 15-point scale of this exam and therefore reflects that the values of total test scores are fairly spread apart. Moreover, due to a sample comprised of beginning through advanced level test-takers, the standard deviation reveals that test-takers were distributed widely across the skill level spectrum. Skewness refers to a distribution of scores that is either negative (left skew) or positive (right skew). Rather than a normal “bell-curved” distribution, a negative skew that resembles a “tail” on the left side of the graph reflects that most of the scores are high, whereas a positive skew reveals that most of the scores within the sample are low. If the skew falls between -2 and 2, the distribution is described as “reasonably normal” (Ockey, 2016). Kurtosis refers to the degree to which the distribution of
scores is either “peaked” or “flattened”. A “perfect” or normal distribution will have a kurtosis of zero. However, in reality, distributions will be either “somewhat” kurtotic or “highly” kurtotic (Ockey, 2016). Pertaining to this exam, the distribution of scores can be described as somewhat negatively skewed at -0.36 with a kurtosis of -0.45. Thus, since both skewness and kurtosis values fall between -2 and 2, it can be said that the distribution is reasonably normal.

Moving on to the test items, I conducted a discrimination analysis including population variance, item variance, item difficulty (IF), item point-biserials, and an item distractor analysis in order to determine how test items performed individually and collectively. Table 2 reflects a population variance of 9.1, a sum of the item variance of 2.914, an alpha score of 0.727, and Standard Error of Measurement score of 1.6.

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L2 Spanish Listening Comprehension Exam discrimination analysis values</strong></td>
</tr>
<tr>
<td>Variance</td>
</tr>
<tr>
<td>9.1</td>
</tr>
</tbody>
</table>

Firstly, population variance and sum of the item variance are calculated and utilized in order to solve for alpha. Alpha is then utilized to calculate Standard Error of Measurement. I will return to an explanation of the alpha score when discussing Research Question 2, therefore I will move on to an explanation of Standard Error of Measurement and an interpretation of the score pertaining to this L2 Spanish Listening Comprehension Exam. Standard Error of Measurement (SEM) is used to determine the reliability of individual test-taker scores on an exam. The SEM score tells us how much “error” is in the individual score. The smaller the SEM score, the smaller the “error” band estimation as to how far the test-taker’s score would vary by chance if
the test-taker hypothetically took the test repeatedly (Carr, 2011). The SEM score for this exam is 1.6. Student 1 scored 6 points out of a total of 15 points. To find the range, we add and subtract the SEM (1.6) to the student’s score (6), which calculates a score range of 4.4-7.6. This means that for this particular test taker, we are 68% sure that their true score is within 1.6 points, plus or minus, their observed score of 6.

There are three statistics used to determine test question difficulty and discrimination for norm-referenced tests. These calculations are referred to as “IF”, “r(p-bis)”, and “ID(UL)”. “IF” stands for item facility or is sometimes referred to as “item difficulty” and reflects the average score on a test item (Carr, 2011). To generate an IF value for an item, the average or mean score is calculated for all responses to the particular item. IF values range from 0.0 to 1.0. For example, a score of 0.0 reflects that zero test-takers answered the item correctly and conversely a score of 1.0 reflects that all test-takers answered the item correctly. However, for norm-referenced tests a range of 0.3-0.7 is considered optimal, with 0.5 considered equal difficulty. Keeping these IF ranges in mind, Table 3 and Table 4 display total number of test items with corresponding item facility scores recorded below each test item. Table 3 displays test items 1 through 8 and Table 4 displays test items 9 through 15.

Table 3

| Item facility scores for test items 1-8 within L2 Spanish Listening Comprehension Exam |
|--------------------------------------|--------------|-------------|-----------------|--------------|-------------|
| Test Item | Q01 | Q02 | Q03 | Q04 | Q05 | Q06 | Q07 | Q08 |
| IF Value | 0.587 | 0.775 | 0.597 | 0.720 | 0.498 | 0.857 | 0.908 | 0.333 |
Although a score of 0.5 is described as reflecting “equal difficulty”, a range of IF’s is preferred to separate the test-takers (Carr, 2011). As seen in both Table 3 and Table 4, items 1-12 and item 15 reflect a healthy range of IF scores between 0.3 and 0.8. The IF values for items 13 and 14 fall below 0.3. An item discrimination analysis will provide further information including insight into why items 13 and 14 initially appear too difficult for this particular test-taker cohort.

Item point-biserials, abbreviated to “r(p-bis)”, is categorized under “item discrimination”, which answers the question “how useful is this item?”, meaning how well does the item separate the test-takers according to ability level. A “good item” will discriminate between high and low ability. The point-biserial is also referred to as the “correlational approach” because the value reflects the correlation of a particular item score with the total exam score (Carr, 2011). A secondary method, the ID(UL) method, also referred to as the subtractive approach, is typically used when the sample size is less than or equal to 15. However, due to a significantly larger sample size, I was able to utilize the correlational approach. Through the correlational approach, a point-biserial range of greater than or equal to 0.30 is considered optimal. To generate an item point-biserial score, a Pearson coefficient is calculated using the scores from the item in correlation to the total exam scores (Carr, 2011). Below, Table 5 and Table 6 display total number of test items with corresponding item point-biserial scores recorded.

<table>
<thead>
<tr>
<th>Test Item</th>
<th>Q09</th>
<th>Q10</th>
<th>Q11</th>
<th>Q12</th>
<th>Q13</th>
<th>Q14</th>
<th>Q15</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF Value</td>
<td>0.655</td>
<td>0.877</td>
<td>0.681</td>
<td>0.534</td>
<td>0.198</td>
<td>0.256</td>
<td>0.650</td>
</tr>
</tbody>
</table>
below each test item. Table 5 displays test items 1 through 8 and Table 6 displays test items 9 through 15.

Table 5

*Point-biserial scores for test items 1-8 within L2 Spanish Listening Comprehension Exam*

<table>
<thead>
<tr>
<th>Test Item</th>
<th>Q01</th>
<th>Q02</th>
<th>Q03</th>
<th>Q04</th>
<th>Q05</th>
<th>Q06</th>
<th>Q07</th>
<th>Q08</th>
</tr>
</thead>
<tbody>
<tr>
<td>r(p-bis)</td>
<td>0.506</td>
<td>0.566</td>
<td>0.464</td>
<td>0.503</td>
<td>0.650</td>
<td>0.295</td>
<td>0.494</td>
<td>0.404</td>
</tr>
</tbody>
</table>

Table 6

*Point-biserial scores for test items 9-15 within L2 Spanish Listening Comprehension Exam*

<table>
<thead>
<tr>
<th>Test Item</th>
<th>Q09</th>
<th>Q10</th>
<th>Q11</th>
<th>Q12</th>
<th>Q13</th>
<th>Q14</th>
<th>Q15</th>
</tr>
</thead>
<tbody>
<tr>
<td>r(p-bis)</td>
<td>0.575</td>
<td>0.504</td>
<td>0.591</td>
<td>0.346</td>
<td>0.352</td>
<td>0.152</td>
<td>0.454</td>
</tr>
</tbody>
</table>

The point-biserial scores for this exam reflect a healthy range of scores between 0.30 - 0.65 for test items 1-13 and item 15. In connection with previous IF values, item 13 reflects an acceptable point-biserial score. Thus, item 13 can be considered difficult but not problematic. However, test item 14 reflects a point-biserial score of 0.152, below the desired minimum score of 0.30. Similar to item 13, when this occurs, it is recommended to additionally consider the item’s IF score. In this case, the IF score for item 14 is 0.256 indicating that item may be too difficult. However, a distractor analysis will provide additional information that can be utilized along with
the point-biserial and IF score in order to determine whether the item should be revised or removed entirely from the exam. Therefore, I will return to item 14 and address the low point-biserial score in connection with results from the forthcoming distractor analysis.

When referencing multiple-choice test items, possible responses are classified as either a “key” denoting the correct response, or a “distractor”, denoting an incorrect response (Brown, 2005). A distractor analysis provides information regarding test-takers’ response frequencies to options for each item as well as the point biserials for each option (key and distractors) per item. This tells us how well or how poorly the distractors are at “enticing” poorly performing students to select them. If a distractor has strong negative correlation (-0.3 or more) with the total test score, it is considered to be a good distractor. If a distractor is close to zero or positive, it is a bad distractor and in need of revision. After running a distractor analysis, test developers can assess which items are in need of revision, and which distractors within an item are successful or problematic for the test’s reliability (Carr, 2011). Table 7 displays the total number of test items with corresponding point-biserial scores for each response. Point-biserial scores in bold typeface reflect that the response score pertains to the key, while point-biserial scores in normal typeface pertain to distractors.
Table 7

Response point-biserial scores for test items within L2 Spanish Listening Comprehension Exam

<table>
<thead>
<tr>
<th>Test Item</th>
<th>a pb(r)</th>
<th>b pb(r)</th>
<th>c pb(r)</th>
<th>d pb(r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q01</td>
<td>-0.357</td>
<td><strong>0.506</strong></td>
<td>-0.118</td>
<td>-0.221</td>
</tr>
<tr>
<td>Q02</td>
<td><strong>0.566</strong></td>
<td>-0.236</td>
<td>-0.340</td>
<td>-0.324</td>
</tr>
<tr>
<td>Q03</td>
<td>-0.224</td>
<td>-0.156</td>
<td><strong>0.464</strong></td>
<td>-0.284</td>
</tr>
<tr>
<td>Q04</td>
<td><strong>0.503</strong></td>
<td>-0.329</td>
<td>-0.199</td>
<td>-0.241</td>
</tr>
<tr>
<td>Q05</td>
<td>-0.235</td>
<td><strong>0.650</strong></td>
<td>-0.437</td>
<td>-0.185</td>
</tr>
<tr>
<td>Q06</td>
<td>-0.234</td>
<td>-0.074</td>
<td>-0.242</td>
<td><strong>0.295</strong></td>
</tr>
<tr>
<td>Q07</td>
<td>-0.271</td>
<td>-0.345</td>
<td><strong>0.494</strong></td>
<td>-0.200</td>
</tr>
<tr>
<td>Q08</td>
<td><strong>0.404</strong></td>
<td>-0.185</td>
<td>-0.217</td>
<td>-0.079</td>
</tr>
<tr>
<td>Q09</td>
<td>-0.372</td>
<td>-0.285</td>
<td><strong>0.575</strong></td>
<td>-0.196</td>
</tr>
<tr>
<td>Q10</td>
<td>-0.221</td>
<td><strong>0.504</strong></td>
<td>-0.291</td>
<td>-0.317</td>
</tr>
<tr>
<td>Q11</td>
<td>-0.320</td>
<td>-0.340</td>
<td>-0.227</td>
<td><strong>0.591</strong></td>
</tr>
<tr>
<td>Q12</td>
<td>-0.099</td>
<td><strong>0.346</strong></td>
<td>-0.180</td>
<td>-0.184</td>
</tr>
<tr>
<td>Q13</td>
<td><strong>0.352</strong></td>
<td>-0.125</td>
<td>-0.295</td>
<td>0.048</td>
</tr>
<tr>
<td>Q14</td>
<td>-0.159</td>
<td>-0.120</td>
<td><strong>0.152</strong></td>
<td>0.066</td>
</tr>
<tr>
<td>Q15</td>
<td>-0.141</td>
<td>-0.273</td>
<td><strong>0.454</strong></td>
<td>-0.298</td>
</tr>
</tbody>
</table>

Focusing on items 1-13 and 15, it is important to note that some distractor values fall just below the optimal value of -0.3. However, considering the healthy range of IF scores and point-biserial scores for these items, we can conclude that these values reflect fairly strong negative correlations in conjunction with the total test score. Returning to item 14, distractor values for response a (-0.159), response b (-0.120) and response d (0.066) reflect a weak negative
correlation to the total test score. Additionally, considering the aforementioned concern related to the item’s low IF value and point-biserial score, a content analysis of the item will provide additional information related to the weak negative correlation.

Figure 6 displays question 14 including item prompt and possible responses. The checkmark next to response “c” indicates the key.

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Según Julia y Marta, durante la Semana Santa las discotecas...</td>
<td>a. se abren en Argentina pero no se abren en España.</td>
</tr>
<tr>
<td></td>
<td>b. se abren sólo el viernes en España.</td>
</tr>
<tr>
<td></td>
<td>c. no se abren en España ni en Argentina. ✓</td>
</tr>
<tr>
<td></td>
<td>d. se abren en España pero no se abren en Argentina.</td>
</tr>
</tbody>
</table>

Figure 6. Screenshot of test item 14. Display of test item 14 prompt, distractors, and key from L2 Spanish Listening Comprehension Exam.

An English translation of the item reads: According to Julia and Marta, during Semana Santa the nightclubs…. (a) are open in Argentina but do not open in Spain, (b) are only open on Friday in Spain, (c) are not open in Spain nor in Argentina, (d) are open in Spain but do not open in Argentina. During the video-recorded discourse, the speakers confer with each other that during the religious holiday of Semana Santa the nightclubs do not open in their respective countries. Therefore, the correct response is (c) are not open in Spain nor in Argentina. A transcript of the recorded conversation is provided below, including an English translation. Note the two differing nouns in bold typeface within the Spanish conversation that translate to “nightclub” in English:
Speaker 1: “Y...y recuerdo ese...eso..esos momentos siempre como de tristeza, ¿no?, donde recuerdo que no se podía ir a la discoteca, por ejemplo...”
(Speaker 1: “And, and I recall that...that occasion always as a sorrowful event, ¿right?, I remember that you couldn’t go out to the nightclubs for example..”)

Speaker 2: (interjects) “¡Claro!”
(Speaker 2: (interjects) “That’s right!”)

Speaker 1: “Y cuando ya era adolecente no podía hacer ese tipo de cosas porque estaba...”
(Speaker 1: “And when I was a teenager I couldn’t do these types of things because it was...”)

Speaker 2: (interjects) “¡Claro! No se abren..los boliches no se abren en Argentina.”
(Speaker 2: (interjects) “That’s right! They don’t open.. the nightclubs don’t open in Argentina.”)

Speaker 1: “En..en España yo creo que era el viernes santo cuando no se abrían.”
(Speaker 1: “In.. in Spain I think it was the Friday of Semana Santa when they wouldn’t open.”)

Speaker 2: “Claro.”
(Speaker 2: “Of course.”)

Although it is not stated verbatim “no se abren en España ni en Argentina” (“are not open in Spain nor in Argentina”), each speaker specifically states that the nightclubs “no se abren” (“they don’t open”). However, while Speaker 1 utilizes “discoteca”, Speaker 2 utilizes “boliche”, which is commonly used colloquially in Argentina and Uruguay in place of “discoteca”. It is plausible that due to limited exposure to the colloquial version of “nightclub” within this particular test-taker population, the majority of test-takers were unable to connect “discoteca” with “boliche” and therefore unable to fully process this portion of the conversation.
Therefore, I identify this item as problematic and advise to remove item 14 from the assessment rather than revise the item.

Connecting the fairly strong negative correlations revealed in the distractor analysis with the exam’s relatively normal distribution of test score and acceptable range of item difficulty, it can be said that an assessment that measures test-takers’ L2 listening comprehension skill based upon video recorded authentic language discourse that incorporates facial cues and body language into the test construct can successfully reflect a test-taker’s L2 listening comprehension skills.

The final component, establishing test reliability leads to Research Question 2: What is the reliability of the test as measured by Cronbach’s alpha? In assessments, when we talk about “reliability” in connection with assessments, reliability refers to the proportion of the test score variance that reflects the “true” score variance, or the test-taker’s “true ability”, minus random error variance (such as wrong answers attributed to the test-taker misunderstanding exam directions). In L2 assessments, we are interested in reliability in order to ascertain how successful an assessment is at reflecting a test-taker’s true or actual ability in relation to test scores. Cronbach’s alpha is utilized in order to measure a test’s reliability (Ockey, 2016). The Cronbach’s alpha for this exam is 0.73. Under Classical test theory (CTT), Cronbach’s alpha is used to describe the estimated true score variance in order to assess for reliability. A test-taker’s score is referred to as the “observed score” that includes the variance of the test-takers “true ability” along with random error variance. Thus, Cronbach’s Alpha reflects how much variance is due to the test-taker’s true ability. Scoring is within the range of 0-1. A score of zero reflects “no reliability” whereas a score of 1 reflects “perfect reliability”. For example, a score of 0.8
conveys that 80% of variance is due to the test-taker’s true ability, while the remaining 20% is due to random error (Carr, 2011). The Cronbach’s alpha formula for reliability is as follows:

\[ \alpha = \frac{k}{k-1} \left(1 - \frac{\sum s^2_i}{s^2_x}\right), \]

where \( k \) is the # of items on the exam, and \( \frac{\sum s^2_i}{s^2_x} \) is the sum of item variance divided by the total score variance. For this exam, \( \alpha = \frac{15}{15-1} \left(1 - \frac{2.914}{9.1}\right) = 0.727 \)

The general rule of thumb is that a score higher than 0.7 reflects relatively good reliability for a particular test population but 0.8 or higher is preferable for high stakes tests (Ockey, 2016). The alpha score for this listening comprehension assessment reveals that nearly 73% variance is due to the test-taker’s true ability within this test population and the remaining 27% is due to random error. This alpha score reflects a low margin of random error, especially considering the limited number of total test items. Thus, the study provides reasonably strong evidence that an L2 Spanish listening comprehension assessment can be developed to the extent that it can be considered a reliable assessment of learner L2 listening comprehension.
CHAPTER 5. SUMMARY AND CONCLUSIONS

To readdress Wagner (2014) concerning the trend of omitting the assessment of L2 listening comprehension due to previous assessment challenges, the results of this study demonstrate that when resources permit, recent technological advances provide the necessary tools to successfully develop and administer reliable and authentic L2 listening comprehension assessments. Furthermore, considering what is at stake for an L2 learner related to assessing and meeting language learner needs, I argue that considering a learner’s L2 listening comprehension in the course placement equation no longer should be “neglected and overlooked” as Wagner (2014) describes. Rather, assessing L2 listening comprehension skills should be viewed as an integral part of the course placement equation.

During the course of the test development and piloting phase of this project there were a few limitations that I will now discuss and provide suggestions for improvement in connection with future test implementation. Firstly, in order to fit the diagnostic and informational session into the 50-minute time slot for each class as well as ensure that test-takers were provided adequate time to take the test, I had to limit the number of videos and total number of test items. Although the exam can be considered a reliable assessment, an addition of 15 items for a total of 30 items could potentially raise the alpha score closer the optimal 0.8 desired for high stakes testing such as a placement exam. For future implementation into a placement exam, I recommend test material that includes a total of 3 videos, approximately 3-4 minutes per video which would provide enough material for 10 test items per video. Taking viewing time into consideration, including listening twice per video, total test duration for the test-taker should be approximately 30-35 minutes to complete. Secondly, utilizing Blackboard to host the exam was
sufficient under the constraints of test delivery in a supervised lab, however was limiting in connection with test format and delivery. Blackboard does not provide the ability to set video viewing parameters, such as limiting playback to one time only. Additionally, only registered ISU students have access to Blackboard and thus only current ISU students could access the exam. For future test implementation that incorporates a listening comprehension portion as part of a Spanish language placement exam, I recommend utilizing a website to host the exam which would permit any potential or incoming new student to take the exam prior to registering for a Spanish course. Each test-taker could be prompted to enter a test code provided by an academic adviser to ensure that only students directly invested in the institution are able to take the exam. After modifying this exam to include a third video and additional test items, future direction for this particular pilot project would be to determine test validity in order to answer the question “how validly have the learner’s been placed?” A preliminary step would be to conduct a descriptive analysis of each course level represented in the sample in order to establish a mean average. The final step would be to establish “benchmarks”, or a range of cut scores to aid the process of test-taker course placement decisions.

In summation, the overarching goal of this research was to determine the extent to which an L2 video-delivered Spanish comprehension assessment can be considered a reliable assessment of L2 listening comprehension. The data analysis reveals a distribution of test-taker scores that parallels the representative distribution of test-taker L2 levels. Therefore, this study suggests that it is possible to create a reliable L2 listening assessment for Spanish that could be utilized to accurately place language learners into level-appropriate courses.
Closing acknowledgements

Before concluding, I would like to once more acknowledge and especially thank Dr. Chad Gasta, Iowa State University’s World Languages and Cultures Dept. Chair, for his tremendous support of this L2 Spanish Listening Comprehension Test Pilot Project. The necessary logistics in order to carry out this research were indeed challenging and entailed reserving the department’s main computer lab for the entire first two weeks of the Fall 2017 semester such that over 400 students would be able to benefit from the experience of taking the diagnostic and learning about how to further develop their L2 listening comprehension skills. When administrators within higher education provide this kind of supportive environment, test developers and instructors are encouraged to push beyond perceived teaching and learning limitations. Through the utilization of available resources coupled with an innovative approach, advancements can be achieved that ultimately foster a positive and impactful learning experience for all students.
REFERENCES


APPENDIX A: PHASE 1 CONSENT FORM

Phase 1 Consent Form

**Title of Study:** Spanish Listening Comprehension Test Development Project

**Investigators:** Jennifer E. Musgrove, TESL/Applied Linguistics graduate student

This form describes a research project. It has information to help you decide whether or not you wish to participate. Research studies include only people who choose to take part—your participation is completely voluntary. Please discuss any questions you have about the study or about this form with the project staff before deciding to participate.

**Introduction**

The purpose of this study is to develop a listening comprehension test for Spanish. This listening comprehension test development project is a preliminary step towards the creation of a listening comprehension Spanish placement exam for the Dept. of WLC in order to provide students with a more comprehensive Spanish placement exam.

You are being invited to participate in this study in order to provide instructor feedback as an integral part of the exam development process.

**Description of Procedures**

If you agree to participate, you will be asked to view the video clips and review the rough draft test items (see attached Word document). After viewing the clips and reviewing the test items please provide feedback sent via email related to anything you regard as important to share, such as audio/video quality, level of interest/engagement of clips, wording of prompts, suggestions for additional prompts, etc..

**Risks or Discomforts**

While participating in this study you may experience the following risks or discomforts: There are no known risks/discomforts.

**Benefits**

If you decide to participate in this study, there will be no direct benefit to you. It is hoped that this research will benefit all potential and current WLC Spanish students by providing students with a more comprehensive Spanish placement exam to ensure proper course placement.

**Costs and Compensation**

You will not have any costs from participating in this study. You will not be compensated for participating in this study.
Participant Rights
Participating in this study is completely voluntary. You may choose not to take part in the study or to stop participating at any time, for any reason, without penalty or negative consequences. If you have any questions about the rights of research subjects or research-related injury, please contact the IRB Administrator, (515) 294-4566, IRB@iastate.edu, or Director, (515) 294-3115.

Confidentiality
Records identifying participants will be kept confidential to the extent permitted by applicable laws and regulations and will not be made publicly available. However, federal government regulatory agencies auditing departments of Iowa State University, and the Institutional Review Board (a committee that reviews and approves human subject research studies) may inspect and/or copy study records for quality assurance and data analysis. These records may contain private information.

To ensure confidentiality to the extent permitted by law, the following measures will be taken: Instructor feedback will be implemented into the exam project without utilizing any identifiable information (such as instructor name).

Questions
You are encouraged to ask questions at any time during this study. For further information about the study, contact Jennifer Musgrove at musgrove@iastate.edu or Dr. Gary Ockey at gockey@iastate.edu

Consent and Authorization Provisions
Your signature indicates that you voluntarily agree to participate in this study, that the study has been explained to you, that you have been given the time to read the document, and that your questions have been satisfactorily answered. Please print a copy of the informed consent for your own files.

Participant’s Name (printed) ________________________________

_________________________________________  ____________________________
Participant’s Signature                           Date
APPENDIX B: PHASE 2 CONSENT FORM

Phase 2 Consent Form for: Spanish listening Comprehension Test Development project

This form describes a research project. It has information to help you decide whether or not you wish to participate. Research studies include only people who choose to take part—your participation is completely voluntary. Please discuss any questions you have about the study or about this form with the project staff before deciding to participate.

Who is conducting this study?
This study is being conducted by Jennifer Musgrove, TESL/Applied Linguistics graduate student.

Why am I invited to participate in this study?
You are being asked to participate because you enrolled in a Spanish class that has you complete a diagnostic Spanish comprehension task. After completing this task, you will receive a diagnostic score. Participating in this research means allowing your answers and the generated diagnostic score to be utilized towards the improvement and development of a future comprehension Spanish placement exam. You should not participate if you are under age 18.

What is the purpose of this study?
The purpose of this study is to develop a listening comprehension test for Spanish. This listening comprehension test development project is a preliminary step towards the creation of a listening comprehension Spanish placement exam for the Dept. of WLC in order to provide students with a more comprehensive Spanish placement exam. Therefore, as explained in the course syllabus:

“*The score generated from this diagnostic exam will not be included in overall course grade components. Additionally, the diagnostic score you receive will not affect your standing in the course or relationship with your instructor or the research investigator*."

What will I be asked to do?
If you agree to participate, you will log into Blackboard and locate the course section provided for this diagnostic exam. Following the instructions provided on the exam, you will be asked to listen to and watch two brief videos in Spanish. You will be instructed to watch each video twice through and then answer 5 multiple choice comprehension questions per video. After completing the test questions you will click “submit” and be proved with the opportunity to review your test score in relation to the correct answers provided on the screen. After all students have completed the exam and reviewed their scores, students will be provided with an informational session related to second language (L2) listening comprehension development including “L2 learner listening tips” that you will be able to use in the classroom and beyond to facilitate the process of listening comprehension.
Your participation will last for approximately 35 minutes.

- 20 minutes: Diagnostic exam including time to review score and correct answers
- 15 minutes: Informational session “L2 learner listening tips for the classroom and beyond”

**What are the possible risks or discomforts and benefits of my participation?**

Risks or Discomforts—There are no foreseeable risks or discomforts to participants.

Benefits—You may benefit from learning more about the significant role listening comprehension plays in language learning, what happens during the listening comprehension process and how you can improve listening comprehension depending upon your current learner level. We hope that this research will benefit all potential and current WLC Spanish students by providing students with a more comprehensive Spanish placement exam to ensure proper course placement.

**What alternatives do I have to participating in the research?**

You may decide to opt out of your generated score from the diagnostic utilized towards future exam improvement and development. If you do not want your diagnostic score utilized in the research please indicate below:

_____ Do not include my score from the diagnostic exam in the statistical analysis for exam improvement and development.

**How will the information I provide be used?**

The information you provide will be used for the following purposes:

The scores generated from this diagnostic exam will be stored in Blackboard to be utilized later towards future exam development and improvement. Diagnostic test scores will not be shared with outside parties. Only the research team will have access to the scores which will be utilized to provide statistical information related to test score data, such as Item Difficulty (IF) and Cronbach’s alpha in order to establish test reliability and provide modification suggestions for potential future implementation of the test.

**What measures will be taken to ensure the confidentiality of the data or to protect my privacy?**

Records identifying participants will be kept confidential to the extent allowed by applicable laws and regulations. Records will not be made publicly available. However, federal government regulatory agencies, auditing departments of Iowa State University, and the ISU Institutional Review Board (a committee that reviews and approves research studies with human subjects) may inspect and/or copy study records for quality assurance and analysis. These records may contain private information.

To ensure confidentiality to the extent permitted by law, the following measures will be taken:

Students’ scores generated from the diagnostic exam will be stored on Blackboard. Each course
will be assigned with a code “A” for beginning level Spanish courses, “B” for intermediate level Spanish courses, and “C” for advanced level Spanish courses. Individual scores will then be assigned a number (1, 2, 3, etc.). An Excel document of scores will be generated using the coded system (A1, A2, A3… B1, B2, B3… C1, C2, C3… ) to be stored on an ISU encrypted secure drive. Coded test scores will be used to gather statistical data such as Item Difficulty (IF) and Cronbach’s alpha in order to establish test reliability. The test and test scores will be stored on Blackboard secure system. Coded test scores on Excel will be stored on an ISU encrypted secure drive. All data will remain stored in Blackboard secure system and ISU encrypted secure drive. A key will not be developed in order to reference back to test participants as this is not necessary for the scope of this project.

Will I incur any costs from participating or will I be compensated?

You will not have any costs from participating in this study. You will not be compensated for participating in this study.

What are my rights as a human research participant?

Participating in this study is completely voluntary. You may choose not to take part in the study or to stop participating at any time, for any reason, without penalty or negative consequences. If you have any questions about the rights of research subjects or research-related injury, please contact the IRB Administrator, (515) 294-4566, IRB@iastate.edu, or Director, (515) 294-3115, Office for Responsible Research, Iowa State University, Ames, Iowa 50011.

Whom can I call if I have questions about the study?

You are encouraged to ask questions at any time during this study. For further information, please contact Jennifer Musgrove at musgrove@iastate.edu or Dr. Gary Ockey at gockey@iastate.edu

Consent and Authorization Provisions

Your signature indicates that you voluntarily agree to participate in this study, that the study has been explained to you, that you have been given the time to read the document and that your questions have been satisfactorily answered. You will receive a copy of the written informed consent prior to your participation in the study.

Participant’s Name (printed) ______________________________________________________________

_____________________________ _________________________
Participant’s Signature Date
APPENDIX C: IRB APPROVAL

From: Office for Responsible Research
Title: Spanish listening comprehension test project
IRB ID: 17-158

Study Review Date: 5/9/2017

The project referenced above has been declared exempt from the requirements of the human subject protections regulations as described in 45 CFR 46.101(b) because it meets the following federal requirements for exemption:

- (1) Research conducted in established or commonly accepted education settings involving normal education practices, such as:
  - Research on regular and special education instructional strategies; or
  - Research on the effectiveness of, or the comparison among, instructional techniques, curricula, or classroom management methods.

The determination of exemption means that:

- You do not need to submit an application for annual continuing review.

- You must carry out the research as described in the IRB application. Review by IRB staff is required prior to implementing modifications that may change the exempt status of the research. In general, review is required for any modifications to the research procedures (e.g., method of data collection, nature or scope of information to be collected, changes in confidentiality measures, etc.), modifications that result in the inclusion of participants from vulnerable populations, and/or any change that may increase the risk or discomfort to participants. Changes to key personnel must also be approved. The purpose of review is to determine if the project still meets the federal criteria for exemption.

Non-exempt research is subject to many regulatory requirements that must be addressed prior to implementation of the study. Conducting non-exempt research without IRB review and approval may constitute non-compliance with federal regulations and/or academic misconduct according to ISU policy.

Detailed information about requirements for submission of modifications can be found on the Exempt Study Modification Form. A Personnel Change Form may be submitted when the only modification involves changes in study staff. If it is determined that exemption is no longer warranted, then an Application for Approval of Research Involving Humans Form will need to be submitted and approved before proceeding with data collection.

Please note that you must submit all research involving human participants for review. Only the IRB or designees may make the determination of exemption, even if you conduct a study in the future that is exactly like this study.

Please be aware that approval from other entities may also be needed. For example, access to data from private records (e.g. student, medical, or employment records, etc.) that are protected by FERPA, HIPAA, or other confidentiality policies requires permission from the holders of those records. Similarly, for research conducted in institutions other than ISU (e.g., schools, other colleges or universities, medical facilities, companies, etc.), investigators must obtain permission from the institution(s) as required by their policies. An IRB determination of exemption in no way implies or guarantees that permission from these other entities will be granted.

Please don't hesitate to contact us if you have questions or concerns at 515-294-4566 or IRB@iastate.edu.