©2021
Nicole Marie Rodríguez
ALL RIGHTS RESERVED
THE PERCEPTION AND PRODUCTION OF LEXICAL STRESS AMONG
EARLY SPANISH-ENGLISH BILINGUAL CHILDREN

By

NICOLE MARIE RODRÍGUEZ

A Dissertation submitted to the
School of Graduate Studies
Rutgers, the State University of New Jersey

In partial fulfillment of the requirements

For the degree of
Doctor of Philosophy
Graduate Program in Spanish

Written under the direction of
Joseph V. Casillas

And approved by

____________________________________

____________________________________

____________________________________

____________________________________

New Brunswick, New Jersey

October 2021
The present dissertation analyzed speech production and perception in early heritage Spanish-English bilingual children. Specically, I investigated how perception and production of lexical stress develops in bilingual children. Current models of second language (L2) speech perception do not include heritage bilinguals or suprasegmental aspects of speech, like word stress. There is nothing inherent about these models that limits their predictions to late L2 learners and segemental aspects of speech, they just have not traditionally been extended to them. By analyzing the development of suprasegemenetal speech perception and production in heritage bilingual children, we can expand the scope of current models of speech perception. To the best of my knowledge, this is the first study to examine how perception and production of
phonology develop in this population of speakers. The project attempts to answer the following questions:

1. How do child bilinguals produce and perceive stress contrasts in both of their languages?

2. How does the production and perception of lexical stress develop as a function of age and proficiency?

3. Is there a discrepancy between perception and production abilities?

4. How does perception and production of stress during childhood compare to that of adult bilinguals?

5. Can lexical stress perception be improved using pedagogical interventions?

6. Does improvement depend on stress pattern and language?

In order to answer these questions, this dissertation is divided into three experiments. Experiment 1 examined the production of lexical stress. Spanish-English early heritage bilinguals ages 6-11 and a group of adult bilinguals completed two production tasks: a Delayed-Repetition task and an Elicited Production task. Data showed that duration is the primary cue that both children and adults use to denote lexical stress in English and Spanish. As a secondary cue, intensity has language- and stress-specific differences for adults and children. F0 was also revealed to be a secondary
cue, but only for the adults. There were no major changes in production strategies based on age or proficiency, indicating that these two factors do not modulate the production of lexical stress in the ages tested.

Experiment 2 explored the perception of lexical stress of English- and Spanish-like pseudowords via an AX discrimination task. The participants were identical to study 1. No effect of language was found, which indicates that participants are able to perceive stress contrasts in English as well they do in Spanish. Adults were shown to be more accurate and more sensitive to stress contrasts than children are. Additionally, age and proficiency were shown to be predictors of sensitivity to stress.

Finally, Experiment 3 investigated the impact of a pedagogical intervention on the perception of lexical stress and the relationship between perception and production abilities. Participants and tasks were similar to Experiments 1 and 2, except the participants underwent a week-long teaching intervention via instructional games that aimed to improve sensitivity to lexical stress. Data showed that sensitivity to lexical stress can be improved after pedagogical interventions. Additionally, this experiment showed relatively similar findings in terms of production in that duration is the main cue that children use to produce stress contrasts.

Overall, the results of the dissertation project inform models of speech learning and pedagogical models to teaching heritage speakers. This dissertation contributes to our understanding of phonological development in bilingual children while also exploring the importance of a teaching intervention to this development. Together,
the three experiments in this dissertation contribute to our understanding of speech learning models in that they support the extension of these models to the perception of suprasegmentals. Furthermore, this dissertation aides in shifting the conversation around heritage speakers away from the focus and comparison between monolinguals and second language learners, to a comparison with adult bilinguals or other heritage language learners.
ACKNOWLEDGEMENTS

First and foremost I would like to thank and express my deepest appreciation to my advisor and committee co-chair, Dr. Joseph Casillas, for his guidance, support, and conversations about Marvel/Star Wars/and everything else through these past 4 years. Thanks for always expecting more of me than I expected of myself. This dissertation would not have been possible without the memes you sent me on Slack. I would also like to extend my gratitude to Dr. Nuria Sagarra, co-chair of this dissertation, for being a source of knowledge and inspiration. Furthermore, I would like to thank Dr. Thomas Stephens for his comments and expertise, which greatly improved my work. Lastly, I am extremely thankful for Dr. Ji-Young Kim for her advice, dedication, and for inspiring this project. I would not have been able to complete this without all of you.

I am grateful to all of those with whom I have had the pleasure to work during these years: my professors in the Department of Spanish and Portuguese and to the staff. I would also like to express my sincere appreciation to the members of RUBilingual and RAP Group Lab (Laura Fernández Arroyo, Kyle Parrish, and Juan José Garrido Pozú). Furthermore, special thanks to all of my colleagues, particularly Julio Caesar Lopez Otero and Benjamin Kinsella. I am very thankful for their support and companionship.
Nobody has been more important to me in the pursuit of this project than my supportive family and friends. Thank you for always being there whenever I need you. A special thanks to my parents, for not kicking me out of the house and doing my laundry for the past 27 years. I truly don’t know where I’d be without you two. Another shout-out to my brother, Alex, who never fails to remind me that I’m not that smart, its truly appreciated. To my friends: Brittanya Salgado, Joey Hanna, Brooke Dobossy, Leah Dobossy, Jenna Hogan, and Kelly Mannion, thank you for staying friends with me even though all I’ve done is complain for the past 4 years. I promise I will be more fun now (even if I am moving across the Atlantic Ocean).

I would also like to thank all of my participants, their families, and all of the schools that supported the completion of this project. Without them, none of this would have happened. I am especially grateful for their participation during a global pandemic and for spending extra time on Zoom even though they were definitely on Zoom overload. I am truly indebted to them.

Finally, thank you to One Direction and my running sneakers. I don’t think I would have been able to finish this dissertation without either of them.
DEDICATION

Para mi sobrina.
TABLE OF CONTENTS

ABSTRACT ................................................................. ii

ACKNOWLEDGEMENTS .............................................. vi

DEDICATION ............................................................ viii

TABLE OF CONTENTS ................................................. ix

LIST OF TABLES ....................................................... xiv

LIST OF FIGURES ...................................................... xv

Chapter 1: Introduction ............................................. 1
  1.1 Introduction .................................................... 1
    1.1.1 Approaches to Speech Perception ...................... 2
    1.1.2 Motor Theory ........................................... 3
    1.1.3 Direct Realist Theory ................................ 7
    1.1.4 General Auditory and Learning Approaches .......... 8
    1.1.5 Speech Perception Models Summary ................... 11
  1.2 L2 Speech Learning Models ................................ 13
    1.2.1 The Perceptual Assimilation Model ................... 14
    1.2.2 The Speech Learning Model ........................... 17
    1.2.3 The Second Language Linguistic Perception Model ... 20
    1.2.4 L2 Speech Model Summary ............................. 23
  1.3 The production and perception interface in L2 speech learning .... 25
  1.4 Pedagogy and phonology .................................... 28
  1.5 Chapter Summary ............................................. 29
  1.6 Organization of the dissertation ......................... 30
Chapter 4: The perception of lexical stress in early heritage bilinguals 77

4.1 Introduction ................................................. 77
  4.1.1 Stress .............................................. 79
  4.1.2 Stress Perception ..................................... 80
  4.1.3 Current Study .......................................... 82
4.2 Method ................................................. 85
  4.2.1 Participants .......................................... 85
  4.2.2 Screening Tasks ....................................... 85
  4.2.3 Experimental Task: AX discrimination .................. 86
  4.2.4 Data analysis ......................................... 88
4.3 Results ................................................ 90
  4.3.1 Response accuracy .................................... 90
  4.3.2 Sensitivity to lexical stress ............................ 95
4.4 Discussion ............................................. 99
4.5 Conclusion ............................................. 101

Chapter 5: Perceptual gains via instructional games: an examination of perception and production of lexical stress in heritage bilingual children ................................................. 103

5.1 Introduction .............................................. 103
  5.1.1 Lexical stress ....................................... 105
  5.1.2 Stress perception and language acquisition .............. 106
  5.1.3 Production of lexical stress ............................. 108
  5.1.4 Perception and production interface ..................... 109
  5.1.5 Bilingual language education ........................... 111
  5.1.6 Current Study ......................................... 113
5.2 Methods ............................................. 115
Appendix G: Posterior predictive checks for statistical models fit to perception data in Chapter 4 ........................................ 164

Appendix H: Posterior predictive checks for statistical models fit to perception data in Chapter 5 ........................................ 166

Appendix I: Power Analysis for AX discrimination - pedagogical intervention ......................................................... 168

REFERENCES ................................................................. 170
### LIST OF TABLES

1.1 Taxonomy of approaches to speech perception. ........................................ 12
2.1 Average of self-reported ratings of skills in both languages ...................... 35
2.2 Countries of origin of the families of all participants. .............................. 36
2.3 Summary of language dominance and proficiency scores for child bilinguals. ................................................................. 37
2.4 Summary of language dominance scores for adult bilinguals. ................. 38
2.5 Example items from the delayed repetition task. ................................... 40
2.6 Trial types from the AX discrimination task. ......................................... 42
3.1 Examples of target items in the delayed repetition task. .......................... 58
3.2 Model output for $\Delta V1-V2$ production metrics in children and adults. 64
3.3 Model output for $\Delta V1-V2$ production metrics and age in children. ...... 67
3.4 Numeric summary of the models examining task effects in children. ....... 68
4.1 Example trial types used in the AX discrimination task. .......................... 87
4.2 Model output of response accuracy as a function of age, proficiency, and language in children. ......................................................... 90
4.3 Numeric summary describing language effect on accuracy scores in adults. 91
4.4 Numeric summary of the parameter estimates of the model examining d’ as a function of age, proficiency, and language. ......................... 95
5.1 Example trial types used in the AX discrimination task. .......................... 119
5.2 Examples of target items in the delayed repetition task. .......................... 120
5.3 Numeric summary of the posterior distributions of the model describing response accuracy. ................................................................. 126
5.4 Model output of change in sensitivity to lexical stress. ............................ 129
5.5 Model output for $\Delta d’$ as a function of age, proficiency, and language. 131
5.6 Numeric summary of the posterior distribution modeling V1-V2 difference as a function of intensity, F0, duration, stress, and language. .... 133
5.7 Model output for V1-V2 production metrics in children. ........................ 134
LIST OF FIGURES

3.1 Forest plot of model estimates for stress correlates. ............... 63
3.2 Parameter estimates as a function of age (children only). ............ 66
3.3 Plot of posterior estimates for production task effects. ............... 70
4.1 Forest plot of model estimates for response accuracy in children. .... 92
4.2 Response accuracy as a function of proficiency, age and language in children. .......................................................... 93
4.3 Three plots showing the effect of language and accuracy scores across languages (children). ................................................. 94
4.4 Forest plot of model estimates for sensitivity to lexical stress. ......... 96
4.5 D’ as a function of proficiency, age, and language. ..................... 97
4.6 Difference in sensitivity to lexical stress between children and adults. 98
5.1 Power analysis for pre/post test discrimination task. ................. 123
5.2 Summary of average correct responses in pre- and post-tests. ........ 126
5.3 Response accuracy as a function of stress patterns. .................... 127
5.4 Change in d’ prime scores between pre- and post-tests. ............... 128
5.5 Δd’ as a function of age and proficiency. ................................. 130
5.6 Forest plot of model estimates for stress correlates. .................. 133
5.7 V1-V2 differences as a function of production metrics, stress, language, and age for each production metric. ......................... 136
D.1 Images used in the elicited production task. ............................. 159
F.1 Posterior predictive checks for production data ....................... 163
G.1 Posterior predictive checks for perception data ....................... 165
H.1 Posterior predictive checks for perception data ....................... 167
Chapter 1: Introduction

1.1 Introduction

Research on heritage speakers and heritage language acquisition has boomed in the last few decades (see Potowski, 2018 for a review). Heritage speakers can be defined as individuals who are “raised in a home where a non-English language is spoken, who speak or merely understand a heritage language and who are to some degree bilingual in English and in the heritage language” (Valdés, 2000). This boom in research on heritage speakers has focused on syntax and morphosyntax, while heritage language phonology has not been extensively studied. The lack of attention to this aspect of heritage bilingualism can be tied to the belief that the heritage phonological system is not different from that of a monolingual in either language. Only in the past decade has heritage language phonology become an area of interest for investigation (Kim, 2015, 2019; Rao & Ronquest, 2015; Ronquest, 2013; Ronquest & Rao, 2018); however, the extant research in this area has centered on adult heritage speakers. This focus on both adult speakers and syntax/morphosyntax has created a gap in our knowledge of how the heritage phonological system develops. This dissertation aims to fill this gap by investigating the speech production and perception of young heritage speakers. Concretely, the present project focuses on the development of lexical stress in early Spanish-English bilinguals.

Lexical stress, as defined by Hualde (2005), is the degree of relative prominence
that a syllable receives compared to others in a word. It represents an ideal case study in Spanish-English bilinguals because it exists in both languages. Crucially, native speakers of English and Spanish process lexical stress differently (Cooper, Cutler, & Wales, 2002; Soto-Faraco, Sebastián-Gallés, & Cutler, 2001), thus allowing for the identification of influence from one language to the other. Although there is a growing body of research that investigates lexical stress in heritage phonology, there has not been a comprehensive study on the acquisition and development of heritage phonology or perception and production of lexical stress in heritage children.

1.1.1 Approaches to Speech Perception

The following section will be broken down into two subsections, the first describes theoretical models and approaches to speech perception and the second considers second language (L2) speech learning models and theories. Both subsections will detail the most important models or theoretical approaches and give empirical evidence in favor of those models. L2 speech models will be described in detail because the present project also considers English, their second language. Furthermore, to the best of my knowledge, there are no models of speech production/perception that specifically address heritage speakers.

The principle motivation behind speech perception research is to describe how spoken language is recognized and understood. In this section I will describe three major theories of speech perception that aim to answer the following question: How
do listeners extract the most fundamental linguistic elements from acoustic signal?

At the core of the issue is the lack of invariance problem, that is, the recognition that there is no simple mapping between units of phonetic structure and units of acoustic structure (Appelbaum, 1996). In other words, there is no one-to-one correlation between the physical properties of speech sounds and the perception of speech sounds. This lack of invariance can make speech perception difficult for monolinguals and bilinguals alike. Nonetheless, listeners come to understand the acoustic signal and interpret meaning from it, despite this invariance. In the following sections the three theories that I will describe are Motor Theory, Direct Realist Theory, and General Auditory and Learning Approaches.

1.1.2 Motor Theory

Originally proposed by Liberman and colleagues in the early 1950s, Motor Theory (MT) claims that the objects of speech perception are articulatory events rather than acoustic or auditory events (Randy L. Diehl, Lotto, & Holt, 2004). This hypothesis means that people’s perception of spoken words occurs through the identification of the vocal tract gestures that are producing the articulatory events rather than by identifying the sound patterns that speech generates. Though MT has undergone changes since its inception (Alvin Meyer Liberman, 1996), every version of the model has claimed that the objects of speech perception are articulatory events rather than acoustic or auditory events. This model states that listeners recover articulatory
events that are actually neuromotor commands to the articulators (known as intended gestures) (Alvin M. Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Following the tenets of MT, early speech perception abilities are indicative of “finely tuned linguistically relevant perceptual abilities” (Miller & Eimas, 1983) or even an “innately given, universal set of phonetic categories” (Eimas & Miller, 1991). An additional claim of MT is that the ability to perceive speech sounds cannot be ascribed to the general mechanisms of audition and perceptual learning, but instead depends on a specialized decoder or module that is speech-specific, unique to humans, and organized and part of the larger biological specialization for language (Alvin Meyer Liberman, 1996; Alvin M. Liberman & Mattingly, 1985). Under this model, the same mechanisms are used for production and perception. The foundation of this theory was built on a series of landmark studies done in Haskins Laboratories that has lead to two major findings. The first finding showed that speech segments typically overlapped one another in time and space (Fowler & Saltzman, 1993; Alvin M. Liberman, Delattre, Cooper, & Gerstman, 1954; Alvin M. Liberman, Delattre, & Cooper, 1952). This coarticulatory relationship made the identification of linguistic units of speech difficult as originally proposed by previous theoretical frameworks.

The second finding is related to the perception of synthetic speech sounds. Haskins Laboratories stipulated that perceived phonemes and features have a simpler relationship to articulation than to acoustics. They also provided a foundation of knowledge of acoustic cues for linguistic units and how complex mapping speech signals to lin-
guistic units is (P. C. Delattre, Liberman, & Cooper, 1955; P. Delattre, Liberman, Cooper, & Gerstman, 1952; Alvin M. Liberman, 1957; Alvin M. Liberman, Delattre, Cooper, & Gerstman, 1954; Alvin M. Liberman, Delattre, Gerstman, & Cooper, 1956; Alvin M. Liberman, Delattre, & Cooper, 1952). An important early discovery in Haskins Laboratories was categorical perception. Typical experiments presented and asked participants to label or identify a series of synthetic consonant-vowel syllables that varied in acoustic parameter and ranged perceptually across several initial consonants. Researchers found two important patterns: first, labeling functions exhibited abrupt boundaries between phoneme categories. This finding means that when there is a gradual change in a variable along a continuum, humans perceive two distinct categories. Second, they found that listeners are able to more successfully distinguish between stimuli that they classified as distinct categories than between stimuli that they had identified as a single category. This finding was an important discovery because it helped to explain the lack of invariance problem to speech perception previously mentioned. If perception the acoustic signal is categorical, it is unlikely that we notice small acoustic differences in the speech signal.

To summarize, this line of research showed that a gradual change in the acoustic signal is not perceived gradually. Specifically, within-category differences seem to be ignored while between-category differences are salient to the listener. Categorical perception likely plays an important role in solving the lack of invariance problem.

Evidence for MT comes, first and foremost, from research conducted at Hask-
ins Laboratories where, as previously mentioned, researchers studied the perception of synthetic speech sounds. The lack of categorical discrimination for certain non-speech sounds supports MT’s theory that speech perception is something unique to humans (Alvin M. Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). However, later experiments with animals showed that categorical perception is not unique to speech sounds, which weakens empirical arguments for MT (Kluender, Diehl, & Killeen, 1987; Kuhl, 1981; Kuhl & Miller, 1975, 1978; Kuhl & Padden, 1982). Note, however, that language experience is also a significant factor in categorical perception, with authors finding evidence of cross-language differences in identification boundaries and discrimination peaks (Abramson & Lisker, 1970; Elman, Diehl, & Buchwald, 1977; Lisker & Abramson, 1970; Williams, 1977). This evidence of speech and nonspeech perception in humans and nonhumans indicates that general auditory mechanisms likely contribute to categorical perception and not specialized mechanisms.

To summarize, there are three main points of this model, the first is that the objects of speech perception are articulatory events rather than acoustic or auditory ones. The second is that perception depends on a specialized decoder that is speech-specific and does not derive from general cognitive mechanisms. This speech-specific device is unique to humans and forms a part of the larger biological specialization for language. Finally, one single mechanism accounts for production and perception. Apart from MT, other theories have since been proposed to explain speech perception, like the
1.1.3 Direct Realist Theory

The next theory, Direct Realist Theory (DRT), like MT, claims that the objects of speech perception are articulatory gestures rather than acoustic events. The central premise of this ecological theory of perception is that the perceiver is directly listening to the perceptual object and not a representation of that object which needs to be inferred or interpreted. This premise means that listeners perceive gestures not through a specialized decoder, but because information in the acoustic signal specifies the gestures that form it. The theory asserts that the objects of perception are actual vocal tract movements, or gestures, and not abstract phonemes or (as in MT) events that are causally antecedent to these movements. This alternative to MT started in the 1980’s and was developed by Carol Fowler (Fowler, 1981, 1986, 1989, 1994, 1996). Furthermore, DRT disagrees with MT in that this model does not believe that specialized mechanisms play a role in speech perception, with the term *direct* implying that perception is not mediated by processes of interference or hypothesis testing. Interestingly, because both MT and DRT claim that the objects of speech perception are gestures, scholars of the two theories often cite the same empirical evidence as support for their claims.

Overall, the DRT is an ecological theory of perception that is similar to MT in that both theories claim that objects of speech perception are gestures. DRT, however,
differs from MT in that those objects of speech perception are considered to be actual, phonetically structured, vocal tract movements and that no specialized mechanisms play a role in speech perception. In order for accurate perception to take place under this model, the listener must recover the gestures to have access to the object being perceived. The two theories presented thus far have posited that objects of speech perception are gestural in nature. They differ in whether or not special mechanisms are invoked for perception. Next, I will summarize the GA framework. Like DRT, GA does not invoke special mechanisms. Unlike both DRT and MT, claims that objects of speech perception are nongestural. GA is a framework and not a theory because it is very abstract in nature and it defines itself by its opposition to the two previously described models.

1.1.4 General Auditory and Learning Approaches

The third and final major approach to speech perception is GA. This framework asserts that general auditory mechanisms are responsible for the similarities in perceptual performance and that there is no need to invoke special mechanisms or modules to explain speech perception. Due to the lack of special mechanisms in this framework, it is hypothesized that speech sounds are perceived using the same mechanisms of audition and perceptual learning that have evolved in humans that handled other classes of environmental noises. Furthermore, perception of gestures does not play a role in the recovery of spoken messages from the acoustic signal. Essentially,
GA proposes that perception does not require any specialized mechanisms and that what listeners are perceiving is derived directly from the acoustic signal, there are no mediating processes or mechanisms.

Lori L. Holt & Lotto (2008) argue that the characteristics of the auditory system would be enhanced by studying it within a more general framework. In doing so, we would improve our understanding of the processing involved in speech perception. Overall, Lori L. Holt & Lotto (2008) have found that there is a disconnect between speech and auditory perception. The authors state that it is necessary to study within a general auditory-cognitive framework because it does not require the dismissal of all processes purported to be specialized for speech signals. It very well could be the case that mechanisms that have evolved or learned processes enacted only on signals that resemble speech play a role in perception (Lori L. Holt & Lotto, 2008).

Evidence to support this framework can be drawn from research on stimulus length effect (Randy L. Diehl & Walsh, 1989) and infant speech data (Lori L. Holt, Lotto, & Diehl, 2003). Here, the perception of the length of an acoustic segment is affected by the duration of adjacent acoustic segments. A target segment will be perceived as shorter next to a long segment than next to a short segment. This applies to both speech and nonspeech sounds. Support for this claim can be found in Randy L. Diehl & Walsh (1989)’s work on the perception of stops and glides. Participants had to categorize speech and non-speech items as having abrupt or gradual onsets. In both speech and non-speech stimuli, a reliable stimulus length effect was observed. There
were no significant differences in labeling between speech and non-speech sounds, which supports the durational contrast account of stimulus length effect. Perceptual compensation for coarticulation under a GA theory would draw from the interactions between stimulus attributes in the auditory system or perceptual learning based on correlated features in the input. Lotto, Kluender, & Holt (1997) was able to show context-dependent responses to consonant-vowel stimuli (/da/ or /ga/) in birds. They posit that the changes in the birds’ responses were not because of the factoring of the signal into gestures but to general auditory interactions between the spectral components of the target and context. Finding this effect in non-human animals goes against the MT argument that certain aspects of speech perception are unique to humans.

Further evidence in support of GA is the fact that, during the early stages of life, infants have the ability to discriminate speech sounds, presumably as a result of their well-developed auditory system that provides sufficient temporal and frequency resolution (Werner & Bargones, 1992). In 1993, (Kuhl, 1993) hypothesized that much of the initial auditory space of the human infant is already segregated by natural boundaries that could underlie many of the speech discrimination results. All in all, the data on infant speech under a GA perspective can be explained as an interaction between the operating characteristics of the auditory system and processes of perceptual learning.

Overall, GA is an approach that does not invoke special mechanisms or modules
to explain speech perception. Here, general auditory mechanisms are responsible for perception. Support for this framework comes from empirical research done with birds and data from infant speech, to name a few.

1.1.5 Speech Perception Models Summary

In this section, I have described MT, DRT, and GA. Each approach attempts to explain how listeners extract the most fundamental linguistic elements from the acoustic signal. MT claims that objects of speech perception are articulatory events and that the ability to perceive speech sounds cannot be ascribed to the general mechanisms of audition and perceptual learning and are actually dependent on a module that is speech-specific. In contrast, DRT claims that perception does not depend on specialized mechanisms and are not mediated by processes of interference or hypothesis testing. Finally, GA also does not have special mechanisms and claims that speech sounds are perceived using the same mechanisms of audition and perceptual learning that have evolved in humans.

Table 1.1, taken from (Randy L. Diehl, Lotto, & Holt, 2004), shows a simplified taxonomy of the major theoretical approaches to speech perception that is based on the classification of special or general mechanisms. The eclectic specialization in the lower left quadrant represents the claim that speech perception uses special mechanisms to recover a nongestural representation of linguistic elements (Randy L. Diehl, Lotto, & Holt, 2004).
Table 1.1: Taxonomy of approaches to speech perception.

<table>
<thead>
<tr>
<th>Special Mechanisms</th>
<th>General Mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gestural</td>
<td>Motor Theory</td>
</tr>
<tr>
<td>Non gestural</td>
<td>Eclectic Specialization</td>
</tr>
<tr>
<td></td>
<td>Direct Realism</td>
</tr>
<tr>
<td></td>
<td>General Approach</td>
</tr>
</tbody>
</table>

At this juncture it is important to point out that these models were designed to describe speech perception at the segmental level. That is, they do not take into consideration suprasegmental aspects of speech that are also used contrastively. While suprasegmentals have not been the focus of early speech perception research, it seems clear that any comprehensive theory of speech perception should encompass suprasegmentals as well. Articulatory gestures related to pitch, for example, could also theoretically be encoded in suprasegmentals and recovered as primitives of perception. Furthermore, the aforementioned models and framework were all created with L1 speech learning in mind. It seems likely that these models cannot thoroughly account for all aspects of bilingual speech, nor should they be expected to, because they were designed with monolingual speech in mind.

The following section will describe the prominent models that have been used in L2 speech learning research. Although heritage speakers are not considered L2 learners, current models do not explicitly account for speech perception in heritage speakers. For this reason I believe it is necessary to detail the models that do exist that take into
consideration bilingual phonology. These models draw from the previously described approaches to speech perception to explain how speech learning may occur over the lifespan.

1.2 L2 Speech Learning Models

Second language speech learning models served as a theoretical foundation to complete this project. Specifically, the three important models in L2 perception that contributed to designing this project were the Perceptual Assimilation Model (PAM), the Speech Learning Model (SLM) and the Second Language Linguistic Perception Model (L2LP). The aforementioned models were developed to explain how speech learning changes over the lifespan with a focus on phonetic segments. The research done under these models, and in speech learning in general, has focused on explaining why “earlier is better” applies to learning L2 phonology. The “earlier is better” argument comes from the Critical Period Hypothesis which posits that the first few years of life are crucial to language development (Lenneberg, 1967). Under this hypothesis, there is an optimal time period for when an individual can best acquire a language. After that period, language acquisition is more difficult and often leads to non-native-like outcomes. Theories have been proposed to explain why foreign accents happen when a language is acquired post-puberty including maturational constraints (McLaughlin, 1977; Sapon, 1952), inadequate phonetic input (Flege, 1992a, 1992b), insufficient motivation (Gardner & Lambert, 1972), or the establishment of incorrect habits in
the early stages of L2 learning (Flege, 1988). Following these assumptions, heritage speakers would have monolingual-like outcomes in their speech perception and production because of how early they started learning both of their languages, though, as we will see, this outcome might not always be the case.

1.2.1 The Perceptual Assimilation Model

The Perceptual Assimilation Model (PAM) is an ecological theoretical perspective to second language speech learning that accounts for issues that are central to how we understand language-specific influences on segmental speech perception and additionally experience-related developmental change. The main goals of this model are to first, provide an account of the types of information that are perceived in speech. Second, the model intends to explain how that information relates to crucial properties of speech production and reveals the phonetic and phonological organization of the listener’s language. Third, the model attempts to explain the ways these factors influence adult’s perception of unfamiliar non-native speech sounds and contrasts. Finally, it explains the development of native and non-native speech (Best, 1995b). Theoretically, PAM makes predictions about how listeners will categorize or assimilate non-native phones and how they will discriminate non-native contrasts. This categorization or assimilation of non-native phones happens with respect to the already existing phonological categories in the native language. This model was founded on a direct realist approach to perception where articulatory gestures are assumed
to be the perceptual primitives for speech perception. It also assumes that native speech is attuned through the discovery of higher-order invariants that specify the gestural constellations that comprise the native phonological inventory (Best, 1995b). Under PAM, perceptual learning of an L2 is determined by nonnative speech perception principles. Perceptual learning is also impacted by age as well as interactions of length of residence, relative usage of L1/L2, and relative quantity and quality of input from native L2 speakers.

In this model, non-native segments are those whose gestural elements or inter-gestural phasing do not match precisely with any already established native segments. Furthermore, non-native segments are usually perceived in accordance with their similarities to, or differences from, native segments that are the closest in proximity to them in the phonological space. Here, patterns are either assimilated to a native category, assimilated as uncategorizable speech sound, or not assimilated to speech and are considered non-speech sounds (Best, 1995b). When listening to an unfamiliar nonnative phone or phonetic segment, naïve listeners are likely to perceptually assimilate the nonnative phone to the most articulatorily similar native phoneme. The process of perceptually assimilating non-native phonemes into our own phonemic inventory places these novel phonemes into one of three possible classifications: (1) categorized exemplar of some native phoneme, where the goodness of fit could be from excellent to poor, (2) uncategorized consonant or vowel that is roughly similar to two or more phonemes so it falls somewhere in between two native phonemes, or
(3) non-assimilable sound that does not have any detectable similarity to any native phonemes.

Best & Tyler (2007) extended PAM to address the issues of L2 learners’ perceptual difficulties and biases across variations in target languages and contrasts. The central question of the revised model, PAM-L2, is “how do nonnative speech perception findings bear on phonological and phonetic aspects of second language perceptual learning?” with the goal of outlining how a common L1-L2 system changes over the course of L2 development (Best & Tyler, 2007, p. 2). There is a frequent assumption that nonnative speech perception can also account for the relative difficulties that late learners have with specific L2 segments and contrasts. PAM diverges from this premise in that both phonetic and phonological levels interact in L2 speech learning, and that this learning depends crucially on the relationship between phonological spaces of the L1 and L2.

PAM claims that L2 listeners classify L2 sound contrasts into different categories depending on the perceived degree of similarity or discrepancy between the native and non-native sounds. How L2 contrasts are classified determines how the contrasts are assimilated to native categories. The classifications are generally based on whether the L2 sounds are perceived as speech or non-speech sounds. If they are perceived as speech sounds, they are further classified into native or non-native (new) categories depending on whether they can assimilate to native categories or not. Non-native contrasts can be classified as Two Category, Category Goodness, or Single Category.
Contrasts that can be discriminated easily are classified as Two Category assimilation, where the two nonnative phones are perceived as acceptable exemplars of two different native phones. For intermediate-level discrimination, Category Goodness predicts that both of the contrasting nonnative phones are heard as tokens of a single native phoneme, but they differ in goodness of fit to that phoneme. Finally, poor discrimination is predicted in Single Category assimilation where two nonnative phones are judged as equally good or poor tokens of the same native phoneme.

To recap, PAM provides a coherent account of the nature of information that is perceived in speech, how that information relates to speech production, and how native and non-native speech develops. Evidence to support this model comes from empirical studies on non-native speech perception in adults and infants. PAM’s claims focus on the naïve listener, which, on the surface, suggests it may not be ideal for modeling heritage phonology given that heritage speakers are not naïve listeners.

1.2.2 The Speech Learning Model

The second model to be discussed is Flege (1995)’s Speech Learning Model (SLM). SLM aims to account for age-related limits on the ability to produce L2 vowels and consonants in a native-like fashion. Work done under this model focuses on bilinguals who have spoken their L2 for many years and not beginning learners because this model is primarily concerned with the ultimate attainment of L2 pronunciation or what more advanced learners are able to do. This model claims that without accurate
perception “targets” to guide the sensorimotor learning of L2 sounds, production of L2 sounds is inaccurate (Flege, 1995). By targets, the model refers to the phonetic input the learners are receiving. This model, however, does not claim that all L2 production errors are perceptually motivated, but many of them have a perceptual bias.

To compare with PAM, PAM agrees with the SLM that the same basic perceptual learning abilities are available to adults learning an L2 as to children learning an L1 or L2. Thus the SLM may provide a framework that is better suited for the current project and it’s focus on children. PAM differs from SLM, however, in that it posits that perceivers extract invariants about articulatory gestures from the speech signal, rather than forming categories from acoustic phonetic cues (Best & Tyler, 2007). Furthermore, PAM’s ecological approach rejects the assumption of mental representation that underlies SLM’s claim that language-specific aspects of speech sounds are specified in long-term memory representations called phonetic categories. Finally, both PAM and SLM agree that L1 and L2 phonological categories exist in a common space. Overall, the two models agree that the same perceptual learning abilities are available to both adult L2 learners and children learning an L1 and that both L1 and L2 phonological categories exist in a common space. They, however, differ on the fact that SLM assumes mental representations and that there is a reliance on the fact that speech sounds are specified as phonetic categories.

Evidence for this model can be seen through the production and perception of
L2 vowels. Phonetic differences between certain L1 and L2 vowels are likely to be discerned by adult L2 learners. When adult L2 learners discern those phonetic differences, new phonetic categories will be established for the L2 vowels. Accordingly, SLM predicts different effects of L2 learning on the production of L2 vowels that is dependent on whether or not a new category has been established for an L2 vowel. Specifically, evidence from the production and perception of vowels (L1 English L2 Spanish: Flege, 1991), categorical discrimination of L1 and L2 vowels (Flege, 1991; Flege & Bohn, 1989b), and the production and perception of initial consonants (Flege & Eefting, 1988; Flege, Munro, & MacKay, 1995; Flege, Takagi, & Mann, 1995).

In short, SLM aims to account for differences in the learnability of phonetic segments in an L2. SLM, in the context of the current project, may be appropriate to analyze heritage phonology. It claims that accuracy of perception of segments in an L2 places an upper limit on the production of those same segments and that the processes and mechanisms that are available to children when acquiring an L1 are also available to adult L2 learners. This comparison of the learning mechanisms being the same for adults and children fits well within the current research. Furthermore, the bidirectional influence between the L1 and L2 phonetic categories can serve to explain the ways in which the heritage and L2 phonetic categories interact with each other in a heritage speaker. Next, the third influential model of L2 perception, the Second Language Linguistic Perception Model, will be described.
1.2.3 The Second Language Linguistic Perception Model

The Second Language Linguistic Perception model (L2LP) is a computational model of acquisition of L2 speech perception and recognition that was originally proposed by Escudero (2005). This model draws on phonetic, phonological, and psycholinguistic constructs to explain L2 learning scenarios.

The model was originally created as a theoretical and methodological framework to describe, explain, and predict the acquisition of L2 sound perception. L2LP is based on the perception component of Functional Phonology (Boersma, 1997). In Functional Phonology, phonological perception grammar handles the mapping of acoustic signal onto discrete segmental units. Like Boersma (1997)’s proposal, L2LP is made up of 3 elements: auditory signal, the device that decodes that signal, and the discrete output of the perceptual mapping. When speech signal is decoded, it leads to the construction of phonological representations, which in Functional Phonology are referred to as perceptual input. The mapping of those speech signals depends on characteristics of listener’s production environment. Linguistic knowledge is assumed to underlie speech perception in this model because speech sound mapping is systematic and language specific in nature. L2LP includes five ingredients that, together, offer a proposal for each of the components of the L2 acquisition process. The ingredients function together to give an explicit prediction, linguistic explanation, and phonetic/phonological description of L2 sound perception at the three logical states.
of the acquisition process. The first ingredient (optimal perception L1 and L2) is proposed to directly enable the explanation of the initial and end states, which are Ingredients 2 and 5. Ingredients 2, 3, and 5 (initial state, learning state, and end state) contain the model’s proposal for the initial, developmental, and end states in L2 sound perception respectively. Ingredient 4 (develop) is the explicit proposal for the L2 learning task and it addresses the L2 development that needs to occur for the attainment of optimal L2 sound perception.

This model attributes difficulties with non-native segments and contrasts to similarities and differences with L1 phonology. It posits that during initial stages of learning, a copy of the L1 perception grammar is made which develops independently from L1 grammar. As learners become more exposed to the target language, they begin to adjust their L2 perception grammar via comparison module. This device is called the Gradual Learning Algorithm which is essentially a general classification device innate in humans. This device allows for the grouping of tokens of objects in the world to optimally cope with all sorts of sensory input. With exposure to the target language, learners eventually learn to selectively activate the L1/L2 perception grammars. Learning is done through 3 learning scenarios: the new scenario, the similar scenario, and the subset scenario. The new scenario is where a contrast is perceived as novel and the learner has to form a new phonetic category. In a similar scenario, contrasts are perceived as familiar, but the learner has to reset the boundary between the acoustic properties of the contrast. Finally, in a subset scenario, a single
category is perceptually assimilated to multiple L1 categories.

Leussen & Escudero (2015) proposed three main revisions to the original L2LP model: (1) the phonologically inspired bias for “faithful” mappings is less restrictive, (2) the possibility of interaction between perception and recognition can be explored, and (3) Jarosz (2013)’s resampling is applied in parsing to enhance the likelihood of convergence. Additionally, their revisions emphasize the fact that the model is meaning-driven, with support from empirical evidence showing that learning correctly predicts the development of L2 phoneme perception.

To empirically test the revisions to the L2LP model, Leussen & Escudero (2015) performed a number of learning simulations to investigate whether the revised model can successfully implement the meaning-driven subset learning scenario described by Escudero (2005). The two-phase experiment (L1 and L2 training) simulates the acquisition of Spanish categories through error-driven learning on lexical items. The results show that learners were able to improve their classification of Spanish minimal lexical pairs (Leussen & Escudero, 2015). The authors state that the results support the current revision and that this revision is more successful in modeling empirical L2 learning results through the implementation of phonetic-phonemic mappings through a more general concept of connection strength. The revision to the L2LP allows for very specific predictions on how L2 experience and L2 input shape the outcome of learning through a workable model of the processes of underlying acquisition of non-native sound systems.
1.2.4 L2 Speech Model Summary

To summarize, PAM is derived from DRT and it proposes that perception of non-native segments is defined by similarities and differences in the proximity of these sounds to native categories. Difficulties with L2 perception in PAM and SLM are based on L1 and target language phonological categories. PAM and L2LP focus on perception of contrasts and how they are assimilated into L1 categories, while the SLM makes predictions about individual segments. Both PAM and SLM claim that L1 and L2 categories share the same phonetic space and that L2 development occurs simultaneously with changes in L1 categories. Unlike PAM and SLM, in the L2LP model, L2 development occurs independently from L1 perception grammar. Here, L2 learners have the ability to develop native-like perception in the L2 without impacting L1 perception grammar.

In the context of the current project it is important to point out that these models do not explicitly account for heritage speakers or suprasegmental aspects of speech production-perception. Due to the fact that L2 speech learning models were created with bilinguals in mind, they are likely better suited for adapting to account for heritage speakers. Additionally, although a priori there is nothing inherent in these models that limits their predictions to segments, they have not generally been used in research on suprasegmental aspects of speech. Essentially, there is a lack of research detailing how these models may account for, or be extended to account for,
speech production-perception beyond the segment. More concretely, few studies, to my knowledge, have tested how the claims of these models can be applied to the learning of lexical stress.

Research into suprasegmentals, like lexical stress, is relevant because they play an important role in language learning, use, and have been shown to influence L2 comprehensibility and accentedness (Anderson-Hsieh, Johnson, & Koehler, 1992; Chun, 2002; Munro & Derwing, 1995, 1998). Trofimovich & Baker (2006) examined the effects of short, medium, and extended L2 language experience on the production of five suprasegmentals (stress timing, peak alignment, speech rate, pause frequency, and pause duration). Adult L1 Korean learners of English completed a delayed sentence-repetition task with declarative sentences. The results suggested that learners with more L2 experience were better able to produce English stress timing like native-speakers of English than learners with less L2 experience. The other metrics associated with production beyond the segmental level (i.e., peak alignment, speech rate, pause frequency, and pause duration) were shown to be related to age of onset of L2 learning. Trofimovich & Baker (2006) asserts that L2 speech-learning models like the SLM and PAM can be extended to account for the results found in their study. The data support the claims of both models because learning was driven by linguistic experience and depended on the segmental and suprasegmental aspect that was being learned.

In sum, despite the fact that the L2 models presented herein focus on segments and
do not explicitly account for lexical stress—the focus of this dissertation—there is no reason to assume they cannot be extended to better describe (1) the distinct types of bilinguals that exist and (2) make predictions about segmental and suprasegmental aspects of phonology. In other words, as with the aforementioned L1 perception models, I believe a complete model of bilingual phonology should encompass not only all aspects of phonology (i.e., segmentals and suprasegmentals), but also all types of bilingualism (i.e., heritage bilingualism).

1.3 The production and perception interface in L2 speech learning

The nature of the relationship between production and perception has been heavily investigated. Successful language learners master both speech production and perception, but the nature of the relationship between the two modalities is unclear. Specifically, it is unknown if this relationship is a causal one, and if it is, which modality causes change and development in the other. It is also a topic of debate as to whether or not a relationship even exists between the two. This relationship can be clearly seen through the correlation between the production of categories and the perceptual judgments of those categories within listeners in monolinguals and bilinguals (Flege, 2003; Flege, Bohn, & Jang, 1997). Theoretically, for monolinguals, speech perception models like MT, DRT, and GA all assume there is a relationship between the two. To begin, MT and DRT believe that this connection between perception and production is established through the role of gestures (whether they are...
intended or actual) as the primitives behind production/perception. MT holds that the perception process is the inverse of production, while DRT affirms that speech perception occurs in combination with all perceptual systems responsible for integrating perceiving and acting (Best, 1995a). On the other hand, GA explains the perception/production relationship through different means because it is not dependent on gestures. One explanation is that perception drives production. This notion can be explained through the general tendency of sound systems to maximize the available phonetic space through a principle of dispersion (Randy L. Diehl, Lotto, & Holt, 2004). There is another account that suggests that production drives perception through the mechanisms of perceptual learning. Here, listeners take statistics on the acoustic correlates available in the speech signal and use them to make inferences about the phonemic properties of incoming speech.

Moving to L2 language learners, the relationship between perception and production seems to be a bit more complex. One explanation of that relationship in L2 learners can be explained in the claims of the SLM. The SLM maintains that L2 phonetic segments are produced as accurately as they are perceived (Flege, 2003). For more concrete evidence that perception drives production, (Flege, Bohn, & Jang, 1997) investigated the vowel production and perception of German, Spanish, Mandarin, and Korean learners of English. The results of this study showed that perceptual accuracy correlated with accuracy producing the same vowel. Additionally, other empirical research has shown that perceptual shifts often lead to gains in production
The opposing view argues that production drives perception (Goto, 1971; Sheldon & Strange, 1982). For support, a semi-longitudinal investigation into the acquisition of the Spanish stop contrast /p/ - /b/ has shown that production changes occurred, and then subsequently perceptual changes occurred (Zampini, 1998). This initial production change is direct evidence that goes against what SLM predicts in that production is the driving force in this relationship. Interestingly though, the participants in this study were advanced speakers and had already differed in their perception of stops at the time of testing when compared to monolinguals. There are, however, other studies that have shown that explicit instruction for L2 production can drive changes in perception (Leather, 1997; Matthews, 1997).

Additionally, there is research that suggests that the relationship between perception and production is not as well defined and clear-cut as MT, DRT, and GA initially proposed. To begin, improvements in perception do not always lead to improvements in production. For support to this claim (Wang, 2002) study of perception and production of American English vowel contrasts in adult L1 Mandarin L2 English learners shows that while the participants improved in their perception of the contrasts, their production did not improve. Furthermore, (Strange, 1995) claims that perceptual changes generally occur more slowly over time when input and use are factored in.

Summarizing, MT, DRT, and GA, all propose a fairly simple relationship between
perception and production. However, the perception-production link seems to be much less clear when taking into consideration bilingual speech, as the extant research provides contradictory evidence.

1.4 Pedagogy and phonology

In addition to contributing to the aforementioned models and further exploring the relationship between speech production and perception, this dissertation also looks at the importance of pedagogy in heritage language development. There is a gap of knowledge surrounding what happens when this group of speakers begins to receive formal education in their heritage language and how that education impacts their phonological development. This project explores the role of pedagogical practices and interventions in the classroom in relation to speech perception. This information on pedagogy is especially relevant because it has been widely accepted that heritage speakers benefit in the area of pronunciation due to their early exposure to the heritage language and do not necessarily need explicit instruction or practice in phonology (Ronquest & Rao, 2018). The findings of this study will lead to practical implications for teaching.

Overall, the present project centers on the acquisition of heritage language phonology in simultaneous bilingual children. Specifically, using a cross-sectional design, I investigate the interaction between two distinct phonological systems, Spanish and English, with a focus on the production and perception of lexical stress. This study
not only provides insight into the development of perception-production abilities in child bilinguals, but also considers the role of bilingual pedagogy in the development of heritage phonology.

1.5 Chapter Summary

This chapter was an introduction to the topic of the present work. Specifically, I provided an overview of the three influential theories related to L1 speech perception (MT, DRT, and GA). L1 speech perception models were designed to describe speech perception at the segmental level and do not take into consideration suprasegmental aspects of speech that are also used contrastively. Additionally, because these models and framework were created with monolingual speech in mind, they cannot be expected to account for a heritage speaker’s speech perception. Although heritage speakers are L1 speakers of their heritage language, L2 speech learning models are better suited to account for their speech learning as L2 models were created with bilingual speech in mind. Next, relevant models of L2 speech learning were discussed along with their similarities and differences. In each section, I detailed the shortcomings and possible applications of these models towards heritage speakers and suprasegmental aspects of speech. Finally, I discussed the relationship between speech perception and production in monolinguals and bilinguals.
1.6 Organization of the dissertation

This work is organized following the format of a three article dissertation. Chapter two is a general methods chapter that details the overarching methodology and procedures implemented in all of the experiments that comprise this dissertation. Chapter three, the first article, explores the production of lexical stress in English and Spanish, as well as how it develops across elementary school ages. Chapter three provides insight into a previously understudied population in relation to how production of lexical stress develops in children, which has been shown to present difficulties for adult second language learners and heritage speakers alike. Chapter four investigates the perception of lexical stress and explores the child sensitivity to minimal stress pairs in comparison with adult bilinguals. Chapter five, the final experimental chapter and third article, investigates the impact of pedagogy in the development of lexical stress, as well as the relationship between perception and production during childhood. Finally, the final chapter of the dissertation discusses the results of the aforementioned experimental chapters, theoretical considerations, and avenues for future research.
Chapter 2: General method

This chapter outlines the general method employed to complete this project. In what follows I provide a description of the questionnaires, the participants who took part in this study, as well as an overview of the experimental tasks, design, and procedure.

2.1 Questionnaires

In order to gather information about the linguistic background of the participant pool, I used of a questionnaire and a phonological assessment to elicit background language information relevant to the research questions (e.g., age of onset of acquisition of Spanish, context of language use, etc.). The questionnaires were the Bilingual Language Profile (BLP) and the Test of Phonological Awareness (TOPA-2+). The results of the BLP and the TOPA-2+ are presented below in Section 2.2.

2.1.1 Bilingual Language Profile

The aim of the BLP is to provide an assessment of language dominance. To this end, the BLP is divided into submodules designed to obtain information about language history, language proficiency, and language use in Spanish and English. The BLP uses Likert-type self-report items, which are weighted by submodule and summed to provide a dominance score ranging from -218 to 218. Scores near either extreme represent dominance in one of the language. The languages are assigned to one of
the extremes arbitrarily. A score near 0 is taken to represent balanced bilingualism. Proficiency, in this dissertation, is operationalized via this language dominance score that includes information and context about language use. It is a measure of how often the participants are using each of their two languages per week, with whom they are using them, self-rating scores, and language attitudes, as opposed to a traditional score on a standardized test that measure syntax and morphosyntactic knowledge. The BLP in its entirety can be found in Appendix A.

2.1.2 Test of Phonological Awareness (TOPA-2+)

After completing the BLP, the children were administered the Test of Phonological Awareness (TOPA-2+). The TOPA-2+ is a norm-referenced measure of phonological awareness for children ages 5 through 8. Norm-referenced tests aim to assign a position to the person being tested in relation to a predefined population. This test measures young children’s ability to (a) isolate individual phonemes in spoken words and (b) understand the relationships between letters and phonemes in English. The test is made up of various components. For the purposes of the present work, I have focused on a final sound identification task and a Letter-Sound test in which students spell simple pseudowords. The schools from which the participants were recruited used this test as an assessment of student progress at the beginning and end of each academic year. In the present study, the scores from the TOPA-2+ are generically referred to as proficiency. Examples of the TOPA-2+ can be seen found in Appendix
2.2 Participants

A total of 77 participants took part in the experiments that make up this dissertation. All participants were Spanish-English bilinguals currently living in the central New Jersey area. All of the children participating in this study were recruited from bilingual schools in central New Jersey. One school in particular, the Greater Brunswick Charter School, served as the main source of participant recruitment. The Greater Brunswick Charter School is a public dual language immersion school servicing students kindergarten through 8th grade. Half of classroom instruction is given in English and the other half is given in Spanish. The overall goal is for students to read, write, listen, and speak the two languages equally well by the end of the program. Using this model of instruction, the school has effectively created an additive bilingual environment that allows students to acquire a second language while maintaining and developing their native language. Though the school is open to students of all language backgrounds, the majority of the student body is made up of Spanish heritage speakers. In the context of this project, heritage speakers are defined as individuals who are raised in an environment where a non-English language is spoken, who speak or understand a heritage language, and who to are bilingual to some degree in English and the heritage language (Valdés, 2000). The participants will have all entered the dual-immersion at the age of five, or Kindergarten. For each language, all participants
provided information regarding the contexts in which they speak, the amount they speak per week, and people with whom they speak. None of the participants knew any languages other than Spanish or English.

2.2.1 Children

In total, 63 children participated in this study. Of those children, 62 children participated in the perception task, only 33 participated in the production task, and 32 participated in the teaching intervention.

They were all consecutive Spanish-English bilinguals who were either born in the United States, or moved to the United States before the age of two and began learning English before the age of five. The students ages ranged from 6;1 to 11;7 years-old. They were grouped by age and also by grade level (number of students per group: Kindergarten, n = 6; 1st, n = 16; 2nd, n = 13; 3rd, n = 12; 4th, n = 5; 5th, n = 11).

Table 2.1 provides a description of the average self-report ratings of how well the children comprehend, speak, read, and write in Spanish and English. In all of the four language skills, the average of the self-reported score was higher in English than it was in Spanish. The biggest differences can be seen in how well they perceive their skills in reading and writing in Spanish in comparison with English. It is interesting to compare with the self-rating of how much they identify as Hispanic/Spanish vs. how much they identify as English/American. For identifying as Hispanic/Spanish, they averaged a score of 4.48, whereas for identifying as English/American the average
score was 4.68.

Table 2.1: Average of self-reported ratings of skills in both languages.

<table>
<thead>
<tr>
<th>Language</th>
<th>Skill</th>
<th>Average Self-Rating (1-6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Comprehension</td>
<td>5.56</td>
</tr>
<tr>
<td></td>
<td>Speaking</td>
<td>5.51</td>
</tr>
<tr>
<td></td>
<td>Reading</td>
<td>4.71</td>
</tr>
<tr>
<td></td>
<td>Writing</td>
<td>4.42</td>
</tr>
<tr>
<td>Spanish</td>
<td>Comprehension</td>
<td>4.73</td>
</tr>
<tr>
<td></td>
<td>Speaking</td>
<td>4.34</td>
</tr>
<tr>
<td></td>
<td>Reading</td>
<td>3.49</td>
</tr>
<tr>
<td></td>
<td>Writing</td>
<td>3.15</td>
</tr>
</tbody>
</table>

The participants self-reported how much they speak English and Spanish with their family and friends. The average percentage of the time that they spent speaking English with their families was 57%, whereas they spent 49% time speaking Spanish with them. With their friends, they speak an average of 75% of time in English and an average of 30% of time in Spanish.

Though all of the students were born in the United States, their families are from Spanish-speaking countries. Table 2.2 provides the country of origin of the family members of the participants and number of participants from that origin.
Table 2.2: Countries of origin of the families of all participants.

<table>
<thead>
<tr>
<th>Country of origin</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Costa Rica</td>
<td>7</td>
</tr>
<tr>
<td>Dominican Republic</td>
<td>13</td>
</tr>
<tr>
<td>Mexico</td>
<td>22</td>
</tr>
<tr>
<td>Puerto Rico</td>
<td>10</td>
</tr>
<tr>
<td>Peru</td>
<td>5</td>
</tr>
<tr>
<td>Spain</td>
<td>6</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>63</td>
</tr>
</tbody>
</table>

Table 2.3 shows mean language dominance scores. Language dominance scores were calculated following the BLP scoring instructions. Each module yielded a particular value of scores per language assessed. To ensure that each module received equal weighing, the total score for each module was multiplied by a predetermined factor to calculate the global language scores. After, the scores per module were added together for each language. To obtain the language dominance index, the total score of Spanish was subtracted from the total for English. A score near zero equates to a balanced bilingual, positive scores indicate dominance in English, and negative scores reflect a dominance in Spanish. The mean, range, and standard deviation scores for TOPA-2+ for the children are also reported in Table 2.3.
Table 2.3: Summary of language dominance and proficiency scores for child bilinguals. Scores derived from the BLP and TOPA-2+.

<table>
<thead>
<tr>
<th>Test</th>
<th>Metric</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLP</td>
<td>Language dominance</td>
<td>61.01</td>
<td>34.82</td>
<td>-13.16 - 143.85</td>
</tr>
<tr>
<td>TOPA-2+ Scores</td>
<td>Proficiency</td>
<td>60.70</td>
<td>16.45</td>
<td>24 - 84</td>
</tr>
</tbody>
</table>

2.2.2 Adults

There were a total of 14 adults who participated in the study. All of the adults were either the parent, guardian, older sibling, or teacher of the children who participated in the study. The adult participants were individuals with whom the children speak to in both of their languages frequently. They were chosen because they are adult bilinguals that provide the input that the children receive. With the idea in mind that the children that participated should be a reflection of the input that they are receiving, which thus creates a more apt comparison group than one with monolinguals or L2 learners.

All adult participants reported using both English and Spanish in their home and work lives, but mainly use English in their professional lives. Adults reported speaking in Spanish with family an average of 70% of the time and with friends an average of 50% of the time. They also reported speaking in English with family an average of 30% of the time and with friends an average of 50% of the time. Two of
the adults were born in Spain and immigrated to New Jersey before the age of 12. One of the adults was born in Puerto Rico and moved to New Jersey before the age of 12. The rest of the adults were born in the central region of New Jersey. All of the adult participants are college educated or are currently finishing their undergraduate degrees. Four of the adults have post-graduate degrees.

The adults can be split into two groups, family members and community members. The family member group comprised 5 parents and 5 siblings. All of the siblings were born in the US and grew up as early sequential bilinguals speaking both English and Spanish with family, friends, and in their communities. None of them received formal instruction in the heritage language. The community member group consisted of 4 teachers that work at the Greater Brunswick Charter School and reported living in the same community as the students. All of the teachers are heritage speakers of Spanish and grew up speaking English and Spanish. One teacher’s family immigrated from Peru, another’s from Puerto Rico, as well as two from Mexico. Due to the fact that the TOPA-2+ was designed for children, it was not apt to measure the phonological awareness abilities of the adult participants in the study. Their BLP scores can be seen in Table 2.4. The adults were mainly English dominant.

Table 2.4: Summary of language dominance scores for adult bilinguals. Scores derived from the BLP.

<table>
<thead>
<tr>
<th>Test Metric</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
</tr>
</thead>
</table>

2.3 Experimental tasks

This dissertation employed the following experimental tasks aimed at measuring production and perception abilities of lexical stress: (1) a delayed-repetition task (DRT), (2) an elicited production task (EPT), and (3) an AX discrimination task.

2.3.1 Delayed repetition production task

The DRT was completed in both Spanish and English. I opted to use a repetition task after considering the ages and reading abilities of the children. It became clear that any production task that was reading based would have caused unwanted problems due to the fact that the children were still learning to read. Participants saw a short phrase on a screen and heard the phrase being read. They were then asked to repeat what they heard (or read). Depending on the reading level of the participants, the phrases were repeated if necessary. There was a total of 32 target sentences and 10 filler sentences. The target phrases contained 3-6 words per phrase with either a past or present tense verb. The target verbs were the second or third word of the utterance, appearing after either a noun phrase or an adverb. The filler sentences were contained first person plural past and present tense verbs. Table 2.5 provides sample target items included in the experiment:
Table 2.5: Example items from the delayed repetition task.

<table>
<thead>
<tr>
<th>Example item</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Ayer <em>cántó</em> el niño.</td>
</tr>
<tr>
<td>b. Yo <em>canto</em> contigo.</td>
</tr>
<tr>
<td>c. I <em>address</em> the crowd.</td>
</tr>
<tr>
<td>d. My <em>address</em> is a secret.</td>
</tr>
</tbody>
</table>

The students were recorded with a Fostex DC-R302 digital recorder and a Shure SM10A dynamic head-mounted microphone. The sound files were segmented and labeled using Praat (Boersma & Weenink, 2021). This production task looked specifically at the verbs in the sentence and compared duration, F0, and intensity of the penultimate (V1) and final vowels (V2) of the minimal pairs. In the case of the above two example sentences, for the verb in sentence 2.5a, V1 is “a” and V2 is “ó.” For the verb in sentence 2.5b, V1 is “a” and V2 is “o.” For a full list of the items in this task, see Appendix C.

### 2.3.2 Elicited production task

The second production task was the EPT, which was only completed in Spanish and by the children. Participants’ utterances were also recorded using a Fostex DC-R302 digital recorder and Shure SM10A dynamic head-mounted microphone and the sound
files were segmented using Praat. The EPT task also looked at duration, F0, and intensity of the penultimate (V1) and final vowels (V2) of the minimal pairs. The purpose of this task was to assess how the children produced Spanish stress contrasts without the presence of a model to follow. Of particular interest was whether or not their realization of stress differed as a function of the task type. Participants were instructed to describe the picture that they were presented as if they were telling a story to a friend. They were shown a series of pictures and their descriptions were recorded. In order to get the contrast in lexical stress, the participants were prompted to describe some pictures both in the past and present with the question/promp:

“Imagine this picture happened yesterday, can you describe this picture like it happened yesterday?” or “Imagine you are the person in this picture, can you describe what you do in the picture?” The inclusion of this task was necessary to see how production changes when participants were given the opportunity to speak more naturally without a model to follow. The first production task, the DRT, was much easier because they were just repeating back what they had heard. The EPT, on the other hand, did not restrict the participants’ responses as much.

2.3.3 AX Discrimination task

After the two production tasks, the participants completed an AX discrimination task that assessed their perception of lexical stress. The experiment was programmed in PsychoPy3 (Peirce et al., 2019) and was hosted online via Pavlovia. The participants
were told that they were going to hear a series of rare words. Their task was to decide if the words they heard were the same or different. They first saw a red cross in the middle of the screen for 150 ms to draw their attention and to signify the start of a new trial. Next, they heard two target words in a sequence. After the presentation of the oral stimuli, the participants saw the words “same” and “different” on the screen. The participants used a button box to record responses. The pairs of words only differed in the placement of lexical stress. For example, they heard pseudoword A followed by pseudoword B and had to decide if the two words were identical. All pseudowords had two syllables and followed the phonotactic rules of either English or Spanish.

There were 40 target trials and 10 distractor trials. The distractor trials contained two completely different pseudowords whereas the target stimuli were presented in one of the four conditions found in Table 2.6:

<table>
<thead>
<tr>
<th>Status</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Same</td>
<td>oxytone - oxytone</td>
</tr>
<tr>
<td>b. Same</td>
<td>paroxytone - paroxytone</td>
</tr>
<tr>
<td>c. Different</td>
<td>oxytone - paroxytone</td>
</tr>
<tr>
<td>d. Different</td>
<td>paroxytone - oxytone</td>
</tr>
</tbody>
</table>

The stimuli were recorded by a 25-year-old female simultaneous bilingual with a Sure SM10A head-mounted microphone and a Fostex DC-R302 digital recorder in a
sound proof booth. Stimuli were randomly presented using PsychoPy3. For a full list of the stimuli used in this experiment, see Appendix E.

2.3.4 Teaching Intervention

Finally, a subset of the participants took part in a teaching intervention through the form of games. For 20 min a day during one week the students played one of the following games: stress bingo, the flyswatter game, and silent ball.

2.3.4.1 Stress Bingo

The game Stress Bingo was played with the purpose of helping the children to recognize syllable prominence within a word. In order to play the game, students had to categorize words based on where the tonic syllable fell within the word. Effectively, this forced the students to pay close attention to where the stressed syllable was in a given word in order to be successful. This game was played through Zoom because some schools were shut down during part of the data collection period due to COVID-19. Originally, the data was collected in Greater Brunswick Charter School, but instruction was moved online starting in March of 2020 and continued online through June of 2021. Thus data collection transferred to Zoom when the instruction format changed to online. For this version of Bingo, there were only two columns that corresponded to either a paroxytone or oxytone stress pattern. They were labeled as Oo or Oo. The students were tasked with classifying the words that the researcher
called out in either column. For example, the researcher would call out “mamá” (Eng. *mom*), and the students would have to put that word in the oO column. The object of the game was to get the most correct responses.

2.3.4.2 Flyswatter game

The flyswatter game is a typical game in the world language classroom. It was chosen because of the students’ familiarity with the format. Here, the focus was also on the identification of the stressed syllable. This game was also played through Zoom with more than one student in the meeting. If there were more than two players, they were divided into two teams, but only two students competed at a time. The two players were unmuted while the rest of the class was left on mute. The researcher said a word and the two students who were competing had to answer as quickly as possible with the answer that matches which syllable is the stressed syllable in the word. For example, the researcher said “cantó” (Eng. *he/she/it sang*), and the students had to yell out “small-BIG.” This game added a time pressure component and a more competitive nature than the previous game because it put two students or teams in competition with each other. Quick, accurate responses were necessary to win the game, which added a degree of difficulty because students did not have extensive time to think about their responses.
2.3.4.3 Silent Ball

This game was the only game played in person before data collection was interrupted. As with the aforementioned games, the goal of this game was to have the participants practice the identification of stressed vs non-stressed syllables. Every student assumed the role of a “weak/small syllable” or a “strong/big syllable.” They had a colored paper taped to their shirt that labeled them weak/small syllables or strong/big syllables. For example blue colored paper indicated weak/small syllables while red indicated strong/big syllables. The researcher said a word out loud and the student with the ball had to say the first syllable in that word and throw the ball to a student that matched the description of the syllable. To illustrate, if the researcher were to say “habló” (Eng. he/she/it spoke), a student could then say “can” (from “cantó,” Eng. he/she/it sang) and subsequently throw the ball to a weak/small syllable student. The student receiving the ball could, in turn, say “tó”—completing the word “cantó”—and then throw it to a “strong/big” syllable student. In essence, the students throw the ball to another student that corresponds with the stress of the syllable they said out loud. If they throw it to the wrong syllable type person, they’re out. This game forced the students to think critically about each syllable in the target word and make decisions based on their knowledge of stress. After a week of game play (totaling 1-2 hours of instruction), the students completed for a second time the AX discrimination task described above.
2.4 Overview of procedures

The participants completed the tasks in the following order: first, they read and signed the informed consent form, or their parent/guardian signed for them. The children who participated were also asked for verbal consent to participate and were told that they were not required to finish the entire project and that they could stop whenever they were tired. Second, participants completed the BLP, children with their parent/guardian, and adults on their own. After the BLP, the children completed the TOPA-2+, which concluded the screening tasks. When the screening tasks were finished, the participants started the experimental tasks. First, they were tasked with the two production experiments, the DRT followed by the EPT. After both production tasks, participants completed the AX discrimination task. Finally, a subset of the participants were chosen to participate in the pedagogical intervention which required roughly an hour of instructional game play spread out over the course of a week. After that week, the participants repeated the AX discrimination task.

2.5 Bayesian Data Analysis

For all statistical inferences I used Bayesian Data Analysis (BDA), which typically took the form of hierarchical regression models. These models provide Bayesian credible intervals, along with other metrics, which I used to quantify uncertainty around parameter estimates. BDA derives a posterior distribution, that is, a distribution of
plausible parameter values, given the data, a data-generating model, and any prior information we have about the parameters being estimated.

The models use Hamiltonian Markov Chain Monte Carlo algorithms to sample values from the posterior distribution. Importantly, this implies that under BDA a single point estimate is not calculated for an effect $\beta$, as is the case with traditional frequentist statistics. Instead, a sample of plausible values for $\beta$ is drawn, which allows us to quantify uncertainty regarding $\beta$ by summarizing the distribution of those values.

In the present work I use four statistics to describe the posterior distribution: (1) the posterior median, (2) the highest density credible interval (HDI), (3) the proportion of the HDI that falls within a Region of Practical Equivalence (ROPE), and (4) the Maximum Probability of Effect (MPE). The median is a point estimate for the middle value of the distribution. The 95% highest density credible interval gives bounds for the effect. The ROPE places a region of practical equivalence around a point null value and calculates the proportion of the HDI that falls within this interval. The MPE calculates the proportion of the posterior distribution that is of the median’s sign, or, in simple terms, the probability that the effect is positive or negative.

To explain with an example: consider a case in which a hypothesis states that $\beta > 0$, and the median point estimate is a positive number, if the 95% HDI of $\beta$ does not contain 0 and is outside the ROPE, and the posterior $P(\beta > 0)$ is close to one, then I
consider there to be compelling evidence for that hypothesis. When taken together, the posterior median, HDI, ROPE, and MPE quantify our uncertainty and provide and intuitive interpretation of any given effect. For a more complete overview of how Bayesian Data Analysis differs from frequentist analyses, see Kimball, Shantz, Eager, & Roy (2016).

2.6 Chapter summary

This chapter provided a detailed outline of the general method used to complete this project. I explained the two questionnaires that were used to assess language background and history information from the participants (BLP) and the proficiency level (TOPA-2+). Next, I described the two groups of participants that took part in the study. After, I went into detail about the 3 experimental tasks (2 production and one perception) and the teaching intervention games that the children took part in. Finally, I discussed an overview of the data collection procedure and the method of data analysis that was chosen. The next chapter is the first of the three articles that make up this project.

It will focus on the two production tasks mentioned in this chapter.
Chapter 3: A cross-sectional look at the production of lexical stress in Spanish-English early bilingual children

3.1 Introduction

Heritage speakers are individuals who are raised in an environment where a non-English language is spoken, who speak or merely understand a heritage language, and who to some degree are bilingual in English and the heritage language (Valdés, 2000). In the case of the current study, the heritage speakers speak English, the majority language, and Spanish, the heritage language. Because these individuals are using and learning a societally dominant language along with a non-dominant language, heritage speakers often exhibit a stronger command of the dominant language. Furthermore, a common claim is that the linguistic skills of these bilinguals in their heritage language often deviate from a monolingual-like norm and are, in some cases, more similar to late second language (L2) learners (S. A. Montrul, 2008; Valdés, 1999). This deviation is often attributed to influence from English and transfer, the process of applying knowledge from one language previously acquired to the target language (Odlin, 1989).

One of the areas in which heritage speakers often deviate from their monolingual counterparts is in speech production. There is a tendency for heritage speakers to have better receptive than productive skills in their heritage language (Au, Oh, Knightly, Jun, & Romo, 2008; Hurtado & Vega, 2004; Kim, 2015, 2019; Knightly, Jun, Oh,
& Au, 2003; Russell N. Campbell, 2000). For example, native speakers of Spanish produce stressed vowels longer than unstressed vowels regardless of the position of lexical stress, whereas heritage speakers have been shown to only produce the stressed vowel longer in words with oxytonic stress (Kim, 2015). Furthermore, Kim (2019) found that heritage speakers did not use the relevant cues to distinguish lexical stress in a consistent manner, resulting in variability in the production of stress.

Previous research on production skills in heritage speakers has focused primarily on adults. The current study focuses on child bilinguals in order to see how production skills develop across the elementary-school ages. Additionally, through studying children, this study can give us insight into how and when the shift to dominance in the majority language (English) influences production. Furthermore, information about that dominance shift can give us information as to how language dominance more generally plays a role in phonetic transfer from the L1 to the L2 as well as from the L2 to the L1.

A possible explanation for better receptive than productive skills in heritage speakers could be use and input. Heritage speakers often only use their heritage language at home with adult members of their household or community and do not always need to respond back in the heritage language (Beaudrie & Ducar, 2005; Hakuta & d’Andrea, 1992; Hurtado & Vega, 2004; Potowski, 2004). This norm of not responding back in the heritage language results in a receptive skills being practiced and used more than productive ones. Participants in the present study were recruited
from bilingual or dual language immersion programs in which they use both Spanish (their heritage language) and English every day in school. This study aims to investigate how production skills related to the realization of lexical stress in both English and Spanish develop in elementary-aged bilinguals. The cross-sectional experimental design can shed light on the developmental process of both languages across the elementary ages. To the best of my knowledge this is the only cross-sectional study that analyzes lexical stress in this population of speakers in this age-group.

3.2 Lexical stress in English and Spanish

Lexical stress can be defined as the degree of relative prominence a syllable receives compared to the others in a word (Hualde, 2005). A prominent syllable is perceived by the listener as being longer, higher, and louder than the others. These impressions are perceptual in nature, but have corresponding acoustic correlates, which are duration, fundamental frequency (pitch), and amplitude (intensity), respectively (Isaacs & Trofimovich, 2012; Ladefoged, 2001; Rogers, 2014). In some languages lexical stress is contrastive, that is, it alone can differentiate meaning between pairs of words. English and Spanish are examples of languages that use lexical stress contrastively, that is, to convey meaning at the word level (Hualde, 2005). Consider the English word pair \textit{PROject} (noun, [ˈprɔˈdʒekt], an individual or collaborative enterprise) and \textit{PROJECT} (verb, [ˈprəˈdʒekt], extend outward beyond something else; protrude), which illustrates the contrastive nature of lexical stress in English.
Words can be categorized as oxytone, paroxytone, or proparoxytone. Oxytones are words where stress falls on the last syllable (e.g., \textit{cantó}, [kan.ˈto], Eng. \textit{he/she/it sang}). Paroxytones are words with stress on the second to last syllable (e.g., \textit{comida}, [ko.ˈmi.ða], Eng. \textit{food}). Finally, in words with proparoxytone stress, stress falls on the third to last syllable (e.g., \textit{clásico}, [ˈkla.si.ko], Eng. \textit{classic}). In Spanish, stress can fall on one of the last three syllables and pitch accented syllables are associated with the same acoustic correlates previously mentioned.

Spanish is traditionally described as a syllable-timed language in which stressed and unstressed syllables have approximately the same duration (Colantoni, Steele, & Escudero, 2015; Hualde, 2005). Nonetheless, duration has been found to be a particularly robust cue for signaling lexical stress in Spanish, along with pitch and intensity (Ortega-Llebaria, 2006; Ortega-Llebaria, Gu, & Fan, 2013; Ortega-Llebaria & Prieto, 2007, 2009). In contrast, English is often described as a stress-timed language in which the intervals between tonic syllables have approximately the same duration. In this language, the absence of stress is typically associated with vowel reduction, unstressed vowels often reduce to the centralized vowel, schwa [o̞], along with secondary cues, i.e., duration, pitch, and intensity. Stressed vowels in English do not reduce and are therefore perceived as being more prominent (Ladefoged, 2001).

Data from perception studies show that English and Spanish speakers process lexical stress differently in their respective first languages (Cooper, Cutler, & Wales, 2002; Soto-Faraco, Sebastián-Gallés, & Cutler, 2001). Speakers of English and Span-
lish have access to the same acoustic cues, but use them in different ways in order to process auditory information. The relative difference in cue weighting between languages could be an explanation as to why second language perception and production of lexical stress in Spanish is difficult for speakers of English.

3.2.1 Production of lexical stress

In children who are learning English as a first language, production of lexical stress is still variable until later on in life (Arciuli & Ballard, 2017; Ballard, Djaja, Arciuli, James, & Doorn, 2012; Hochberg, 1988; Kehoe, Stoel-Gammon, & Buder, 1995; Pollock, Brammer, & Hageman, 1993; Schwartz, Petinou, Goffman, Lazowski, & Cartusciello, 1996; Vihtnan, DePaolis, & Davis, 1998). Typical patterns of production of lexical stress in young English-speaking children can include the deletion of a weak syllable if it precedes a strong syllable. (G. Allen, 1978; G. D. Allen & Hawkins, 1980; Echols & Newport, 1992; Gerken, 1994). Additionally, production of trochaic words with stress correlates in an adult-like fashion can be seen around age 3 (Ballard, Djaja, Arciuli, James, & Doorn, 2012). Ballard, Djaja, Arciuli, James, & Doorn (2012) found that children ages 3-7 years old were not completely adult-like in terms of their use of intensity when producing words that start with a weak syllable (e.g., “tomato”), which have a less frequent stress pattern in English. Older children, around the ages of 8-11 years old, produced lexical stress in an adult-like way, but are not quite on par in terms of intensity contrasts for words that start with a weak syllable (Arciuli
As for Spanish monolinguals, some variability is also present in stress production throughout childhood (Hochberg, 1988). In a study of 50 Mexican-American preschoolers of ages 3-5, Hochberg (1988) had the children produce nonce words in stress minimal pairs. The results of the experiment revealed that in general nonce words that followed stress rules of Spanish were produced better than nonce words that did not follow them. When comparing the two age groups, Hochberg (1988) found that the 5-year-olds produced fewer errors than the 3-year-olds. However, the older children’s error rates were still relatively high when it came to producing irregular stress. A more comprehensive study with a wider age-range of children is needed for generalizations to be made about whether Spanish-learning children produce stress correlates in an adult-like manner.

Few studies have investigated production of lexical stress in heritage speakers. Those that have focused on adults Kim (2015). For instance, Elias, McKinnon, & Milla-Muñoz (2017) studied the role of code-switching and lexical stress on Spanish vowel production. They found that the participants reduce the quality and shorten the duration of the unstressed syllables. In a similarly vein, Ronquest (2016) examined Spanish vowel production in different speech styles. Both Ronquest (2016) and Elias, McKinnon, & Milla-Muñoz (2017) found that these speakers used duration and vowel quality to mark lexical stress in Spanish. Additionally, both studies found that stressed vowels were produced with longer duration when compared to
unstressed vowels. Vowel reduction to denote the unstressed syllable is not characteristic of Spanish monolingual speech. These two studies suggest that heritage speaker production of lexical stress in Spanish can be influenced by English phonology.

Kim (2015) examined of the production of lexical stress in adult English dominant heritage speakers. She found that in nearly half the cases, heritage speakers produced unstressed vowels that were longer than the stressed vowels in paroxytones. Furthermore, Kim (2019) found that heritage speakers deviated from the monolinguals and performed similarly to the L2 learners when producing Spanish stress minimal pairs.

Taken together, the few studies that have investigated lexical stress in heritage speakers show that there is a tendency for their production strategies to deviate from monolinguals of the heritage language. Much is still left to be discovered about the production patterns of heritage speakers, especially in terms of how they develop across childhood.

3.2.2 The present study

The present study aims to explore the production of lexical stress in Spanish and English across the elementary school ages in Spanish-English bilingual children, and to compare their speech to that of adult bilinguals. Heritage speakers tend to have better receptive than productive skills in their heritage language (Au, Oh, Knightly, Jun, & Romo, 2008; Hurtado & Vega, 2004; Kim, 2015, 2019; Knightly, Jun, Oh, & Au, 2003; Oh, Jun, Knightly, & Au, 2003; Russell N. Campbell, 2000). Previous
research into lexical stress has focused on adult L2 learners and heritage speakers (Beaudrie, 2017; Guion, Harada, & Clark, 2004; Kim, 2015, 2019; Li & Grigos, 2018; Romanelli, Menegotto, & Smyth, 2015; Saha & Mandal, 2016; Zhang, Nissen, & Francis, 2008). This semi-longitudinal and cross-sectional study of the development of lexical stress in heritage speakers allows us to explore how the child’s languages influence each other while developing across the elementary school ages.

The present study addresses the following research questions:

1. How do child bilinguals produce stress contrasts in both of their languages?
2. How does the production of lexical stress develop as a function of age?
3. Does the production of lexical stress differ depending on the task type?

Following previous research, I hypothesize that the child bilinguals will produce stress contrasts differently when compared with the adult bilinguals. Furthermore, I expect that age should mediate stress production as it has been shown that a shift to dominance in English happens as heritage speakers get older. Finally, with respect to the third research question, I expect task type to have an impact on stress production. Specifically, I believe that when participants have a model to follow they will be more likely to produce stress contrasts in a target-like manner.
3.3 Method

3.3.1 Participants

A total of 39 participants completed the production tasks. There were two groups, a group of children and a group of bilingual adults (n = 6), who were the parents or guardians of the children participating. The adult participants served as a comparison group. There were a total of 33 children (ages 6-9.6 years old) that participated in the study. All of the children were recruited from dual language immersion schools in the central New Jersey area and are all consecutive bilinguals, having acquired Spanish first and then English before the age of five. The children were measured on their phonological abilities through a standardized test and completed the Bilingual Language Profile (BLP) in order to assess language history, proficiency, and use. None of the participants knew any languages other than Spanish or English at the time of testing.

3.3.2 Experimental Tasks

This study included two production tasks aimed at eliciting lexical stress in two different contexts. The first is a Delayed Repetition Task (DRT) and the second is an Elicited Production Task (EPT).
3.3.2.1 Delayed repetition

The DRT was broken down into two parts: one part in English and the other in Spanish. In this task, participants were simultaneously presented with auditory and visual stimuli. The visual stimuli was a short phrase on a screen and the auditory stimuli was the same phrase being read out loud. After the sentence finished playing, the participants repeated what they heard/saw. The sentence did not disappear from view and they could refer back to it as needed. There was a total of 32 sentences. Each sentence contained 3-6 words with either a target verb in past or present tense. Target words never came first or last in a sentence. The following are samples of the task in both English and Spanish.

Table 3.1: Examples of target items in the delayed repetition task.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Ayer <em>(cantó)</em> el niño.</td>
</tr>
<tr>
<td>b</td>
<td>Yo <em>(canto)</em> contigo.</td>
</tr>
<tr>
<td>c</td>
<td>I <em>address</em> the crowd.</td>
</tr>
<tr>
<td>d</td>
<td>My <em>address</em> is a secret.</td>
</tr>
</tbody>
</table>

Participants were recorded with a Fostex DC-R302 digital recorder with a Shure SM10A dynamic head-mounted microphone. The sound files were segmented and labeled using PRAAT.
3.3.2.2 Elicited Production Task

The second production task was conducted only in Spanish. It was also recorded with a Fostex DC-R302 digital recorder and the sound files were segmented using PRAAT. The EPT aimed to get a more natural production of lexical stress in Spanish as compared with the previously described DRT. Participants were presented with a series of pictures. They were told that they needed to describe the pictures as if they were telling a story to a friend. Their descriptions of the pictures were recorded. The instructions were given in Spanish. In order to assure that target items were produced with the desired stress contrasts, the participants were prompted to describe some pictures both in the past and present with the question/prompt: “Imagine this picture happened yesterday, can you describe what is happening in this picture like it happened yesterday?” or “Imagine you are the person in this picture, can you tell me what you do in this picture?” The order of the past and present prompts were counter-balanced.

3.3.3 Acoustic measures

Synchronized waveform and spectrographic displays were used to analyse and segment the recordings and mark the onset and offset of each vowel within the same word (e.g., /a/ and /o/ in hablo and habló). Recordings were first screened to ensure that no abrupt pauses were found before or after the target words. A Praat (Boersma &
Weenink, 2021) script extracted duration (ms), F0 (Hz), and intensity (dB) values at the mid-point. The V1-V2 difference (henceforth \( \Delta V_1 - V_2 \)) for each word for each metric was calculated by taking the V1 value and subtracting from it the V2 value. For example, in the word “hablo” (Eng. *I speak*), the duration in milliseconds of V2 was subtracted from the duration of V1 in order to calculate the V1-V2 difference (e.g., V1: 134 ms; V2: 92 ms; \( \Delta V_1 - V_2 = 42 \)). In order to make the three metrics (duration, F0, intensity) comparable, the values from each were standardized by subtracting each individual value from the metric mean and dividing by the standard deviation.

### 3.3.4 Data analysis

I report five primary statistical analyses with the same basic structure in order to address the aforementioned research questions. The standardized values for each metric were analyzed using Bayesian multilevel regression models. The models examined how the production metrics varied as a function of a series of predictors (described below). In all cases, the data generating process was assumed to be distributed as normal. All analyses were conducted in R (R Core Team, 2019). The models were fit using Stan (Stan Development Team, 2018) via the R package *brms* (Bürkner, 2017). Prior beliefs about the parameters being estimated were included using regularizing, weakly informative priors (Gelman, Simpson, & Betancourt, 2017). The priors were normally distributed and centered at 0 with a standard deviation of 0.2
for all population-level parameters. The only exceptions were the model intercept \(N(0, 2)\) and the correlation of grouping effects \(\text{lkj}(2)\). Models were fit with 2000 iterations (1000 warm-up). Hamiltonian Monte-Carlo sampling was carried out with 4 chains distributed between 4 processing cores, thus the posterior distributions for each parameter estimated consisted of 4,000 samples that were compatible with the data and our prior beliefs.

For statistical inferences I describe in detail the posterior probability distributions, for each parameter of interest as well as the 95% highest density interval (HDI) and the maximum probability of effect (MPE). Furthermore, I established a region of practical equivalence (ROPE) of \(\pm 0.1\) and, consequently, report the percentage of the 95% HDI contained within the ROPE. The goodness-of-fit of all models was assessed using posterior predictive checks (see Appendix F).

3.4 Results

3.4.1 Child and adult production of lexical stress

The first series of analyses modeled the standardized values of the V1-V2 differences of each production metric (duration, F0, intensity) as a function of lexical stress (paroxytone, oxytone) and language (Spanish, English). The metric, level of stress, and level of language were coded into an indicator variable and the omnibus model analyzed \(\Delta V1 - V2\) for each indicator. The model did not include a population
intercept, but there were grouping factors for participants and word items. The
effects of stress and language varied for each participant. Figure 3.1 plots the posterior
distributions of $\Delta V1 - V2$ by stress type and language for each metric, in children
and adults.

Focusing on duration, one can clearly observe that this metric was the primary
cue for realizing stress differences in both children and adults, as the posterior distri-
butions were outside of the established ROPEs by a substantial margin.

F0, for its part, was mainly used by adults as a secondary cue for indicating stress.
In particular, the adults used pitch to mark oxytonic stress in English ($\beta = 0.5$, HDI
$= [0.07, 0.92]$, ROPE = 0.01, MPE = 0.99) and paroxytonic stress in Spanish ($\beta =$
-0.78, HDI = [-1.16, -0.37], ROPE = 0, MPE = 1). The children only came close to
realizing a stress difference via F0 in Spanish paroxytones ($\beta = -0.17$, HDI = [-0.35,
-0.02], ROPE = 0.16, MPE = 0.98), but approximately 16% of the HDI fell within
the ROPE.

Intensity, on the other hand, appears to be a secondary cue used more by children
than adults. Specifically, in English, children produced a substantial V1-V2 difference
for paroxytones ($\beta = 0.77$, HDI = [0.61, 0.93], ROPE = 0, MPE = 1), but not oxytones
($\beta = -0.06$, HDI = [-0.25, 0.12], ROPE = 0.65, MPE = 0.74), and in Spanish they
produced a V1-V2 difference for oxytones ($\beta = -0.42$, HDI = [-0.58, -0.26], ROPE =
0, MPE = 1), but not paroxytones ($\beta = -0.12$, HDI = [-0.28, 0.04], ROPE = 0.41,
MPE = 0.93). The effect was much less stable in adults in both languages for both
Figure 3.1: Summary of posterior model estimates of $\Delta V_1 - V_2$ as a function of stress and language for each production metric (duration, F0, intensity). Points represent posterior means and horizontal bars indicate the ±66% and 95% highest density credible intervals. Colors indicate stress type and point shapes represent to two groups. The vertical discontinuous line represents the point of null value (0) surrounded by a region of practical equivalence (ROPE) of ±0.1.
paroxytone and oxytone words, as large margins of the posterior probability mass fell within the predetermined threshold of practical equivalence. The complete model output is summarized in Table 3.2.

Table 3.2: Numeric summary of the posterior distribution modeling $\Delta V1-V2$ as a function of language and stress for each production metric. The table includes summaries for both children and adults.

<table>
<thead>
<tr>
<th>Language</th>
<th>Metric</th>
<th>Stress</th>
<th>Group</th>
<th>Median</th>
<th>95% CrI</th>
<th>MPE</th>
<th>%ROPE</th>
<th>Rhat</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Dur.</td>
<td>Ox.</td>
<td>Adults</td>
<td>-0.59</td>
<td>[-0.89, -0.28]</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3875.27</td>
</tr>
<tr>
<td></td>
<td>Dur.</td>
<td>Ox.</td>
<td>Children</td>
<td>-0.49</td>
<td>[-0.63, -0.33]</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1538.79</td>
</tr>
<tr>
<td></td>
<td>Dur.</td>
<td>Parox.</td>
<td>Adults</td>
<td>0.95</td>
<td>[0.59, 1.31]</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3242.57</td>
</tr>
<tr>
<td></td>
<td>Dur.</td>
<td>Parox.</td>
<td>Children</td>
<td>0.94</td>
<td>[0.78, 1.1]</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1530.2</td>
</tr>
<tr>
<td></td>
<td>F0</td>
<td>Ox.</td>
<td>Adults</td>
<td>0.5</td>
<td>[0.07, 0.92]</td>
<td>0.99</td>
<td>0.01</td>
<td>1</td>
<td>3129.13</td>
</tr>
<tr>
<td></td>
<td>F0</td>
<td>Ox.</td>
<td>Children</td>
<td>0.16</td>
<td>[0, 0.31]</td>
<td>0.98</td>
<td>0.24</td>
<td>1</td>
<td>1621.85</td>
</tr>
<tr>
<td></td>
<td>F0</td>
<td>Parox.</td>
<td>Adults</td>
<td>0.08</td>
<td>[-0.32, 0.51]</td>
<td>0.65</td>
<td>0.37</td>
<td>1</td>
<td>2323.58</td>
</tr>
<tr>
<td></td>
<td>F0</td>
<td>Parox.</td>
<td>Children</td>
<td>0.11</td>
<td>[-0.09, 0.29]</td>
<td>0.88</td>
<td>0.45</td>
<td>1</td>
<td>1786.19</td>
</tr>
<tr>
<td></td>
<td>Int.</td>
<td>Ox.</td>
<td>Adults</td>
<td>-0.1</td>
<td>[-0.58, 0.41]</td>
<td>0.66</td>
<td>0.32</td>
<td>1</td>
<td>2392.15</td>
</tr>
<tr>
<td></td>
<td>Int.</td>
<td>Ox.</td>
<td>Children</td>
<td>-0.06</td>
<td>[-0.25, 0.12]</td>
<td>0.74</td>
<td>0.65</td>
<td>1</td>
<td>1439.1</td>
</tr>
<tr>
<td></td>
<td>Int.</td>
<td>Parox.</td>
<td>Adults</td>
<td>0.4</td>
<td>[-0.18, 0.91]</td>
<td>0.94</td>
<td>0.08</td>
<td>1</td>
<td>2041.4</td>
</tr>
<tr>
<td></td>
<td>Int.</td>
<td>Parox.</td>
<td>Children</td>
<td>0.77</td>
<td>[0.61, 0.93]</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1592.81</td>
</tr>
<tr>
<td>Spanish</td>
<td>Dur.</td>
<td>Ox.</td>
<td>Adults</td>
<td>-0.6</td>
<td>[-0.95, -0.29]</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2342.52</td>
</tr>
<tr>
<td></td>
<td>Dur.</td>
<td>Ox.</td>
<td>Children</td>
<td>-0.56</td>
<td>[-0.71, -0.39]</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1634.99</td>
</tr>
<tr>
<td></td>
<td>Dur.</td>
<td>Parox.</td>
<td>Adults</td>
<td>0.46</td>
<td>[0.12, 0.82]</td>
<td>0.99</td>
<td>0</td>
<td>1</td>
<td>3168</td>
</tr>
<tr>
<td></td>
<td>Dur.</td>
<td>Parox.</td>
<td>Children</td>
<td>0.36</td>
<td>[0.2, 0.51]</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1463.86</td>
</tr>
<tr>
<td></td>
<td>F0</td>
<td>Ox.</td>
<td>Adults</td>
<td>0.18</td>
<td>[-0.15, 0.57]</td>
<td>0.85</td>
<td>0.27</td>
<td>1</td>
<td>2127.62</td>
</tr>
<tr>
<td></td>
<td>F0</td>
<td>Ox.</td>
<td>Children</td>
<td>-0.04</td>
<td>[-0.21, 0.11]</td>
<td>0.68</td>
<td>0.8</td>
<td>1</td>
<td>1540.23</td>
</tr>
</tbody>
</table>
To summarize, the analyses show that, in this data, duration is the primary cue for indicating lexical stress in Spanish and English for children and adults. That being said, differences do exist in the way children and adults use F0 and intensity. Specifically, adults tended to use F0 more, while children made use of intensity.

### 3.4.2 Lexical stress and age

The next analysis fit the same model as described in the previous section with the addition of an age and age $\times$ indicator predictors. Age was standardized and included in the model as a continuous predictor. This model was only fit to the child data.

Figure 3.2 plots 150 draws from the posterior along with the line of best fit for $\Delta V1-V2$ as a function of age in standardized units for each language and each metric. Qualitatively, it appears that the use of intensity, F0, and duration remains relatively constant across the ages that were tested in this experiment. The model summary available in Table 3.3 corroborates this affirmation, as there were no reliable interactions between age and any of the metrics that I used to measure lexical stress.
Figure 3.2: The posterior distributions of each parameter from a model that was refit to only include children. V1-V2 differences of intensity, duration, and F0 across both stress types and both languages are shown as a function of age. Age was standardized.
Table 3.3: Numeric summary of the model examining age effects in children. The table summarizes the posterior distribution modeling $\Delta V1-V2$ as a function of language, stress, and age for each production metric.

<table>
<thead>
<tr>
<th>Language</th>
<th>Metric</th>
<th>Stress</th>
<th>Median</th>
<th>95% CrI</th>
<th>MPE</th>
<th>%ROPE</th>
<th>Rhat</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Dur. x z-Age</td>
<td>Ox.</td>
<td>-0.04</td>
<td>[-0.15, 0.06]</td>
<td>0.76</td>
<td>0.89</td>
<td>1</td>
<td>1224.23</td>
</tr>
<tr>
<td></td>
<td>Dur. x z-Age</td>
<td>Parox.</td>
<td>0.09</td>
<td>[-0.05, 0.24]</td>
<td>0.89</td>
<td>0.54</td>
<td>1</td>
<td>1829.52</td>
</tr>
<tr>
<td></td>
<td>F0 x z-Age</td>
<td>Ox.</td>
<td>0.02</td>
<td>[-0.13, 0.18]</td>
<td>0.58</td>
<td>0.82</td>
<td>1</td>
<td>2071.97</td>
</tr>
<tr>
<td></td>
<td>F0 x z-Age</td>
<td>Parox.</td>
<td>0.19</td>
<td>[0.02, 0.36]</td>
<td>0.99</td>
<td>0.13</td>
<td>1</td>
<td>1888.15</td>
</tr>
<tr>
<td></td>
<td>Int. x z-Age</td>
<td>Ox.</td>
<td>0.12</td>
<td>[-0.07, 0.3]</td>
<td>0.9</td>
<td>0.43</td>
<td>1</td>
<td>1828.39</td>
</tr>
<tr>
<td></td>
<td>Int. x z-Age</td>
<td>Parox.</td>
<td>0.14</td>
<td>[-0.01, 0.29]</td>
<td>0.96</td>
<td>0.31</td>
<td>1</td>
<td>1863.41</td>
</tr>
</tbody>
</table>

| Spanish  | Dur. x z-Age | Ox.  | -0.05  | [-0.21, 0.1]  | 0.74 | 0.74  | 1    | 1905.46 |
|          | Dur. x z-Age | Parox.| 0.12   | [-0.03, 0.27] | 0.93 | 0.4   | 1    | 1985.99 |
|          | F0 x z-Age  | Ox.  | 0.02   | [-0.13, 0.17] | 0.61 | 0.82  | 1    | 1885.77 |
|          | F0 x z-Age  | Parox.| 0.03   | [-0.12, 0.21] | 0.66 | 0.76  | 1    | 2124.07 |
|          | Int. x z-Age| Ox.  | -0.11  | [-0.26, 0.04] | 0.92 | 0.45  | 1    | 1684.27 |
|          | Int. x z-Age| Parox.| 0.05   | [-0.1, 0.2]   | 0.71 | 0.76  | 1    | 1948.88 |

3.4.3 Task effects

The final two models addressed the question of whether or not the realization of lexical stress was affected by the formality of the task in children. The first model fit the standardized $\Delta V1-V2$ metric as a function of task. The task predictor was sum coded (DRT = -1, EPT = 1) such that the model estimated the main effect and the
model intercept represented the grand mean. The model included grouping variables for participants and items, and the effect of task varied by participants. Given the model, the data, and our prior beliefs, there was no evidence suggesting that the experimental task had an effect on $\Delta V1-V2$ ($\beta = 0.01$, HDI = $[-0.03, 0.05]$, ROPE = 1, MPE = 0.66). Next, I included the indicator variable representing all combinations of lexical stress and language for each metric. The indicator $\times$ task interaction slope was allowed to vary for each participant. None of the resulting posterior distributions provided compelling evidence for a task effect. The full summary of each model is presented in Table 3.4 and illustrated in Figure 3.3.

Table 3.4: Numeric summary of the models examining task effects in children.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Median</th>
<th>95% CrI</th>
<th>MPE</th>
<th>%ROPE</th>
<th>Rhat</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DV ~ task</td>
<td>Intercept</td>
<td>0.04</td>
<td>[-0.08, 0.15]</td>
<td>0.76</td>
<td>0.87</td>
<td>1</td>
<td>1120.27</td>
</tr>
<tr>
<td></td>
<td>Task effect</td>
<td>0.01</td>
<td>[-0.03, 0.05]</td>
<td>0.66</td>
<td>1</td>
<td>1</td>
<td>5453.95</td>
</tr>
<tr>
<td>DV ~ Task * metric</td>
<td>Ox. En. Dur.</td>
<td>-0.09</td>
<td>[-1.22, 1.11]</td>
<td>0.56</td>
<td>0.13</td>
<td>1.01</td>
<td>803.06</td>
</tr>
<tr>
<td></td>
<td>Ox. En. Int.</td>
<td>0.12</td>
<td>[-2.82, 2.75]</td>
<td>0.53</td>
<td>0.06</td>
<td>1</td>
<td>3674.39</td>
</tr>
<tr>
<td></td>
<td>Ox. En. F0</td>
<td>-0.03</td>
<td>[-2.8, 2.88]</td>
<td>0.51</td>
<td>0.06</td>
<td>1</td>
<td>3578.98</td>
</tr>
<tr>
<td></td>
<td>Ox. Sp. Dur.</td>
<td>0.07</td>
<td>[-1.14, 1.2]</td>
<td>0.54</td>
<td>0.13</td>
<td>1.01</td>
<td>808.31</td>
</tr>
<tr>
<td></td>
<td>Ox. Sp. Int.</td>
<td>0.13</td>
<td>[-1.08, 1.25]</td>
<td>0.58</td>
<td>0.13</td>
<td>1.01</td>
<td>810.71</td>
</tr>
<tr>
<td></td>
<td>Ox. Sp. F0</td>
<td>0.06</td>
<td>[-1.14, 1.19]</td>
<td>0.54</td>
<td>0.13</td>
<td>1.01</td>
<td>802.63</td>
</tr>
<tr>
<td></td>
<td>Parox. En. Dur.</td>
<td>-0.47</td>
<td>[-3.16, 2.6]</td>
<td>0.62</td>
<td>0.05</td>
<td>1</td>
<td>3929.18</td>
</tr>
<tr>
<td></td>
<td>Parox. En. Int.</td>
<td>-0.29</td>
<td>[-3.22, 2.47]</td>
<td>0.58</td>
<td>0.05</td>
<td>1</td>
<td>4324.62</td>
</tr>
</tbody>
</table>
3.5 Discussion

The present study investigated the production of lexical stress in Spanish-English heritage bilingual children. A group of bilingual adults served as a comparison group. I administered two production tasks, a DRT and an EPT with the goal of eliciting lexical stress contrast in minimal pairs.

I aimed to answer three research questions with this study. The first question addressed how child bilinguals produce stress contrasts in both of their languages. The results show that in both languages, duration is the primary cue for signaling stress. In comparison with their adult participants, the child production appears to align perfectly, as they are also using duration to signal stress. This finding corroborates previous research showing that duration is a primary cue for producing stress contrasts in Spanish (Elias, McKinnon, & Milla-Muñoz, 2017; Ronquest, 2016). Apart from duration, F0 is also used as a secondary cue, albeit in a language-specific manner. That is to say, adult bilinguals use pitch in Spanish paroxytones and English oxytones. The child bilinguals, on the other hand, did not use pitch in the same manner. Unlike the adults, the children also consistently used intensity to mark stress contrasts in
Figure 3.3: Posterior distributions for task effect models. Panel A plots the joint probability distribution of the EPT and DRT tasks. The blue point represents the posterior median for each distribution. Panel B represents the posterior distribution of the task effect. Panel C illustrates the posterior distributions of the task effect as a function of language and stress for each production metric. For panels B and C, the white points represent the posterior means ±66% and 95% credible intervals.
English paroxytones and Spanish oxytones. The use of F0 in adults could be due to the target word positions, as all target words fall in non-final positions which would elicit prenuclear pitch stress. If word position were to impact F0 cue use, then we should expect to see the same F0 cue effect in children because all participants were presented with the same stimuli in the same conditions.

The latter two findings align with previous studies that show variability present in stress production throughout childhood (Hochberg, 1988). Although there were cue-weighting differences, for the most part, the children were making use of the primary cue, duration, similarly to the adults. In contrast to the findings of Kim (2015), where English dominant heritage speakers produced unstressed vowels longer than stressed vowels in words with paroxytonic syllable stress, no such difference was found in the present analysis. There were language specific differences found for both English and Spanish paroxytones, but neither involving duration. This difference could be due to the fact that the heritage speakers in this study attended a dual immersion school. They had approximately the same amount of input and instruction in both languages, so they likely should not be considered English dominant like the participants in Kim (2015). It is important to note that the results of the BLP suggest that they are English dominant, but this measure includes self-reporting and self-ratings which could lend to subtle differences in what encompasses the meaning of dominance.

Focusing solely on English, English paroxytones show V1-V2 differences in intensity. This difference aligns with previous work, where children were not completely
adult-like in their production of stress. Though perhaps any comparison between the two studies should be taken with caution because Arciuli & Ballard (2017) and Ballard, Djaja, Arciuli, James, & Doorn (2012) used polysyllabic words instead of bisyllabic words like the current study.

The second research question was concerned with how the production of lexical stress develops during early childhood. In the data analyzed, age did not modulate the cues used for stress production. This finding is interesting because at first glance does not seem to align with previous research like Hochberg (1988) where age did modulate production. It should be noted, however, that in the aforementioned study the participants were tasked with saying pseudowords and in the current study they repeated real words. Additionally, the children in the current study were slightly older than the oldest participants in Hochberg (1988) (ages 6-9,6 years old as compared to 3;0-5;11). It could be that by the time children reach 6 years-old (the youngest age tested in the present study), they are close enough to adults in terms of cue use, so that age does not end up being an important predictor for modulating cue use. The lack of an effect of age in this study as compared to Hochberg (1988) could be due to the fact that real words are easier to produce than pseudowords are.

There is ample evidence for variability in production, and general differential production when heritage speakers are compared to a monolingual-like norm (Au, Oh, Knightly, Jun, & Romo, 2008; C. B. Chang & Yao, 2016; C. Chang, Haynes, Yao, & Rhodes, 2008; Hurtado & Vega, 2004; Kim, 2015, 2019; Knightly, Jun, Oh, & Au,
The differential use of cues as compared to adult bilinguals could be attributed to two main reasons. First, as previously mentioned, the comparison of heritage bilinguals to monolinguals or second language learners that is the norm in previous research could be the main reason that there is well documented evidence of production differences in the literature and only slight differences in the current study. Due to the fact that there is a more even comparison between groups in terms of language experience, the minimal differences could reflect that comparison. Second, the heritage speakers in this study were receiving formal education in their heritage language at the time of testing, which was not always the case for the participants that have been studied previously. Lack of formal education in the heritage language, as related to quality and quantity of input, has often been attributed to the monolingual-heritage differences found in previous studies (S. Montrul, Foote, & Perpiñán, 2008).

The final question addressed whether production of lexical stress was affected by the task type. To this end I compared the data from the DRT with that of the EPT. On the one hand, the DRT provides more experimental control and lends itself to testing participants at virtually any age. On the other hand, the EPT provides a more naturalistic context that is less likely to be affected by “lab speech.” Only the children participated in both tasks and there was no evidence suggesting that the experimental task affected the use of intensity, F0, or duration. The lack of a task effect found between a naturalistic measure of speech and one where participants were
asked to directly repeat a model speaker leads us to believe that the EPT was not more difficult for the participants. The inclusion of the second task makes it possible to affirm that the participants could produce the stress contrasts without a model to follow. They effectively did so when they were given the freedom to come up with their own responses.

In general, the study aimed to frame the discussion of heritage language production a bit differently than previous research in which there have been comparisons made to monolinguals of the heritage language or L2 speakers. In contrast to previous work, the exploratory nature of this project hopes to solely discuss the capabilities of the children participating in the study. The comparison group chosen reflected the bilingual input that the children received in their daily lives (parents, siblings, and teachers). Here, the use of other bilingual community members makes for a better comparison of language abilities than a comparison with monolingual speakers of the heritage language or second language learners who have an entirely different experience with learning the language.

For studies that include older heritage speakers, the comparison group could still come from the speaker community that the heritage speakers belong to, but it instead could be made up of the older bilingual members who immigrated to the U.S. or of first generation citizens who grew up speaking the heritage language at home and learned English in school, then went on to raise bilingual children. Shifting the focus away from a comparison with monolinguals of the heritage language to focusing solely on
the way in which the heritage language is being produced in comparison with similar members of the speech community is vital because its changes the conversation from one of a deficit to a focus on what bilinguals with a similar linguistic experience can tell us about the bilingual experience.

Using adult speakers in their community allows for a better comparison because they make up the input that the children being studied are receiving. The group of adult bilinguals was used in order to have an adequate comparison with an attainable standard instead of a comparison with a monolingual speaker, as bilinguals are not two monolinguals in one (Grosjean, 1989).

3.6 Conclusion

This chapter detailed the results of the first experiment that investigated how the production skills of bilingual heritage speakers develop during the early stages of formal education. Specifically, I analyzed the production of lexical stress in English and Spanish. The results revealed that duration is the primary cue that both children and adults use to produce stress contrasts. As a secondary cue, adults bilinguals use F0 in a language-specific manner (for Spanish paroxytones and English oxytones). In comparison, children used intensity as a secondary in a language-specific manner (for Spanish oxytones and English paroxytones). In contrast to previous research, age and proficiency did not play a role in mediating stress production, as the children used all three metrics to measure stress here in a fairly consistent manner across the age
range tested in the current experiment. Additionally, the formality of the task did not impact stress production. Finally, the results motivate the use of bilinguals with similar linguistic experiences as comparison groups for research with heritage speakers in order to gain insight into bilingual phonological development and the heritage bilingual experience more generally. The next chapter will delve into the AX discrimination task that aimed to explore the perception of lexical stress in heritage bilinguals.
Chapter 4: The perception of lexical stress in early heritage bilinguals

4.1 Introduction

Second language (L2) speech learning models aim to describe how spoken language is recognized and understood. Research under L2 models such as the Perceptual Assimilation Model (PAM) (Best, 1995b), Speech Learning Model (SLM) (Flege, 1995), and Second Language Linguistic Perception model (L2LP) (Escudero, 2005) has focused on primarily on segments and not on suprasegmental aspects of speech. Examining the acquisition of suprasegmentals, like stress, is important to L2 acquisition because it has been shown to influence L2 comprehensibility and accentedness (Anderson-Hsieh, Johnson, & Koehler, 1992; Chun, 2002; Munro & Derwing, 1995, 1998). Few studies have attempted to extend the claims of the aforementioned models to suprasegmentals. Trofimovich & Baker (2006), for instance, asserts that L2 speech-learning models like the Speech Learning Model (SLM) and Perceptual Assimilation Model (PAM) can be extended to account some suprasegmental measures, e.g., stress timing, peak alignment, speech rate, pause frequency, and pause duration.

Apart from not explicitly accounting for suprasegmental aspects of speech, first (L1) or L2 speech perception models do not take into consideration heritage speakers, as they do not fit into a model of monolingual language acquisition and they are not late L2 learners. Heritage speakers are early bilinguals that grow up speaking the heritage language (the minority language) at home and the societally dominant
language at school (Valdés, 1999). Previous research with this population has shown that they oftentimes exhibit better comprehension than production abilities. For example, Oh, Jun, Knightly, & Au (2003) studied Korean heritage speakers and L2 learners’ perception and production of dento-alveolar Korean stop consonants. The results showed that the heritage speakers of Korean outperformed the L2 learners on measures of perception, but not on the production tasks.

The current project studies lexical stress perception. Previous research into lexical stress has focused on adult L2 learners and heritage speakers (Beaudrie, 2017; Guion, Harada, & Clark, 2004; Kim, 2015, 2019; Li & Grigos, 2018; Romanelli, Menegotto, & Smyth, 2015; Saha & Mandal, 2016; Zhang, Nissen, & Francis, 2008). This work presents a cross-sectional study of the development lexical stress in heritage speakers. The main purpose of this research is to shed light on how both languages influence each other during acquisition across the elementary school ages. To this end, it explores the bidirectional relationship between the L1 and L2 phonology. This study is novel in that it is one of the first, to the best of my knowledge, that looks at the perception abilities of heritage speaker children in lexical stress during the early stages of development. It adds to our knowledge of early heritage bilinguals and models of speech perception.
4.1.1 Stress

The current project investigates the perception of lexical stress. Following the definition provided by Hualde (2005), stress is the degree of relative prominence that a syllable receives compared to others in a word. For example, when considering the words “pinto” (Eng. *I paint*) and “pintó” (Eng. *he/she/it painted*), the stressed syllable in the first word is the first syllable (“pin”) whereas the stressed syllable in the second word is the second one (“tó”). Stress was chosen for this research because similarities can be drawn about the way stress functions in both Spanish and English.

In both languages, stress in phonologically contrastive, that is, it has a distinctive function. For example, the difference between the words “papá” (Eng. *dad*) and “papa” (Eng. *potato*) can only be found in the position of the stressed syllable. “Papá” is an oxytonic word because the stress falls on the last syllable whereas “papa” is a paroxytonic word with the stressed syllable falling on the second to last syllable. Furthermore, the two languages differentiate stress by pitch, duration, and intensity. Overall, stress functions similarly, but not identically, in both languages. For this reason, lexical stress, as a suprasegmental feature of English and Spanish, lends itself to experimental research on bilingualism because it provides a window into how exactly the two languages might influence each other.
4.1.2 Stress Perception

Research on L1 perception of stress has shown that stress patterns are acquired early on (Pons & Bosch, 2010). Factors such as syllable weight, morphological category, subregularities in the lexicon, and the influence of segmentally similar words play a role in the perception of word stress (Timothy Lee Face, 2000; Timothy L. Face, 2006; Pons & Bosch, 2010). Previous research has demonstrated that English and Spanish speakers process lexical stress differently in their respective L1s (Cooper, Cutler, & Wales, 2002; Soto-Faraco, Sebastián-Gallés, & Cutler, 2001). Evidence from Cooper, Cutler, & Wales (2002) and Soto-Faraco, Sebastián-Gallés, & Cutler (2001) suggests that English and Spanish listeners use suprasegmental cues in different ways when processing auditory information. This differential use of cues could be an explanation as to why L2 perception of stress in Spanish is difficult for L1 speakers of English.

The manner in which the perception of speech sounds develops and is acquired in child heritage bilinguals has been largely under-researched. In contrast, the myriad of research done with monolingual infants shows that during the first year of life babies lose the ability to discern speech sounds from all languages and become more sensitive to their first language speech contrasts. This ability applies to segmental Werker & Tees (1984) and suprasegmental levels of speech Skoruppa et al. (2013).

Despite an abundance of evidence showing that children are able to discern speech sounds in their L1 from a very young age, research on L2 learners and heritage
speakers alike points to some degree of difficulty with the perception speech sounds in the L2 and heritage language, respectively. These difficulties in L2 speakers have been attributed to the L1 prosodic system and early exposure (Chrabaszcz, Winn, Lin, & Idsardi, 2014; Lee, Shin, & Garcia, 2019; L1: Li & Grigos, 2018; Ortega-Llebaria, Gu, & Fan, 2013; early exposure: Dupoux, Peperkamp, & Sebastián-Gallés, 2010; Guion, Harada, & Clark, 2004). Previous research with heritage speakers has yielded mixed results, some show minimal or no perception differences (Boomershine, 2013; Kim, 2015; Ronquest, 2013; Ronquest & Rao, 2018; Willis, 2005) while others show differences in perception due to less experience with Spanish and dominance in English (Alvord, 2010b, 2010a; Henriksen, 2012; Rao & Ronquest, 2015).

Focusing only on the perception of lexical stress, studies have shown that English-learning infants show a preference for the predominant stress pattern in their native language (trochaic stress) as early as 9 months old Turk, Jusczyk, & Gerken (1995). Spanish-learning infants are able to discriminate between trochaic and iambic stress, but do not show a trochaic bias like English-learning infants do Skoruppa et al. (2009). Pons & Bosch (2010) suggests that this could be due to the fact that there are many words in Spanish-learning infants’ input that contain iambic stress (e.g., bebé ‘baby,’ mamá ‘mommy,’ papá ‘daddy’).

To the best of my knowledge, no studies have focused on how child heritage bilinguals perceive lexical stress in Spanish. However, adult heritage bilinguals’ perception of lexical stress has been studied. Two studies have explored how heritage speakers
and L2 learners discriminate stress minimal pairs in Spanish (Kim, 2015, 2019). In an identification task, heritage speakers performed on par with a monolingual comparison group when discerning contrasts between minimal pairs (Kim, 2015). In the same study, L2 learners had significantly lower accuracy rates on the perception task, which alludes to the notion that early exposure is beneficial for the perception of lexical stress. Further evidence for this advantage can be seen in Kim (2019), which investigated the perception of stress minimal pairs in various prosodic contexts. The results of the perception task suggested that the heritage speakers were as successful as the monolinguals in identifying the location of stress. As in, Kim (2015), the heritage speakers and monolinguals showed significantly higher accuracy rates in the perception task than the L2 learners.

4.1.3 Current Study

The present study aims to explore the development of the perception of lexical stress in Spanish and English in bilingual children. L2 speech learning models do not explicitly account for the acquisition of suprasegmentals. Their predictions have been previously extended to account for suprasegmentals, like stress. One study in particular, Trofimovich & Baker (2006), examined effects of short, medium, and extended L2 language experience on the production of five suprasegmental measures (stress timing, peak alignment, speech rate, pause frequency, and pause duration). In this example adult L1 Korean learners of English completed a delayed sentence-repetition
task with declarative sentences. The results revealed that learners with more L2 experience were better able to produce English stress timing like native-speakers of English than learners with less L2 experience, leading Trofimovich & Baker (2006) to conclude that their findings support the notion that L2 speech-learning models could be extended to include suprasegmentals.

The present study will expand on that research. Lexical stress was chosen because it has been shown to be persistently difficult for L2 and heritage speakers to perceive and produce in a monolingual-like way (Beaudrie, 2012, 2017; Carreira, 2002; Rao & Kuder, 2016). Previous research into lexical stress has focused on adult L2 learners and heritage speakers (Beaudrie, 2017; Guion, Harada, & Clark, 2004; Kim, 2015, 2019; Li & Grigos, 2018; Romanelli, Menegotto, & Smyth, 2015; Saha & Mandal, 2016; Zhang, Nissen, & Francis, 2008). This cross-sectional study of the development lexical stress in heritage speakers allows for the exploration of how both languages influence each other during acquisition across the elementary school ages. It explores the bidirectional relationship between the L1 and L2 phonology. This study is novel in that it is one of the first, to the best of my knowledge, that looks at the perception abilities of heritage speaker children in lexical stress across their development.

The following questions guide the study:

1. Do child Spanish-English heritage speakers perceive stress contrasts in both of their languages?
2. Do proficiency and age mediate sensitivity to stress?

3. How does perception of stress during childhood compare to that of adult bilinguals?

The aforementioned questions aim to help us gain a better understanding of the development of early bilingual children’s perception of lexical stress. With the previously reviewed literature in mind, I hypothesize that the Spanish-English heritage speakers will perceive stress contrasts well in both of their languages. This prediction is motivated by previous research showing that adult heritage speakers perceive lexical stress well in comparison to monolinguals (Kim, 2015, 2019). Furthermore, I expect proficiency and age to mediate sensitivity to stress where older and higher proficiency speakers will be more sensitive to stress due to the fact that difficulties in differences in perception have been previously attributed to less experience (proficiency) with Spanish and dominance in English (Alvord, 2010b, 2010a; Henriksen, 2012; Rao & Ronquest, 2015). Finally, I expect that, in comparison, the children will perceive stress just as well as the adult bilinguals do, as L1 stress patterns are acquired early on (Pons & Bosch, 2010).
4.2 Method

4.2.1 Participants

In total, 77 participants completed the experiment, 62 child Spanish heritage speakers (ages 6-11, 7 years old) and a comparison group of 15 adult bilinguals (their parents or other adult speakers in their lives such as teachers and older siblings). All of the children that participated in this study are consecutive bilinguals that were recruited from dual language immersion schools servicing students in Kindergarten through 8th grade in the central New Jersey area. Participants were measured on their phonological abilities through a standardized test that the schools have adopted to track student progress in their program. Additionally, they completed the Bilingual Language Profile (BLP) in order to assess language history, proficiency, and use. None of the participants have knowledge of any languages other than Spanish or English.

4.2.2 Screening Tasks

In order to ensure that the participants were all evenly matched in terms of input, exposure, and proficiency, they completed the BLP and a proficiency task. The BLP was filled out by the parent/guardian of the student along with the informed consent form. The BLP contained questions about the participants’ experience and exposure to language, including age of exposure to Spanish and English, number of years in formal education in each language, location and length of time living abroad. There
was also a Likert scale skills assessment of their Spanish and English abilities in reading, writing, speaking, and listening. The goal was to ensure homogeneity among the participants and to control for possible confounds.

As a proxy for proficiency, the students completed an assessment of their phonological abilities. The assessment is one that the school has implemented and uses as a means of tailoring instruction to meet the needs of the students. The test used is the Test of Phonological Awareness (TOPA-2+), which is a measure of phonological awareness for children ages 5 through 8 years. This test measures young children’s ability to (a) isolate individual phonemes in spoken words and (b) understand the relationships between letters and phonemes in English. The test has various components, but of central importance to the present work are the final sound identification task and the Letter-Sound test. The school district adopted this test in order to keep track of student progress.

4.2.3 Experimental Task: AX discrimination

The experimental task aimed to gather information on the participants perception of lexical stress. The perception task was an AX discrimination task that assessed the participants abilities to discern differences between oxytonic and paroxytonic stress. The experiment was programmed through PsychoPy3 software (Peirce et al., 2019). In this task, the participants were told that they were going to hear a series of rare words and that they needed to decide if the pairs of words they heard were exactly the
same or different. A red cross appeared in the middle of the screen for 150 ms to draw their attention and to signify the start of a new trial. After the red cross disappeared, they heard the two target words in a sequence. Once the oral stimuli stopped playing, the participants were presented with the words “same” and “different” on the screen. The participants used a button box to record their responses. The stimuli were presented in one of either four conditions:

Table 4.1: Example trial types used in the AX discrimination task.

<table>
<thead>
<tr>
<th>Status</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Same</td>
<td>oxytone - oxytone</td>
</tr>
<tr>
<td>b. Same</td>
<td>paroxytone - paroxytone</td>
</tr>
<tr>
<td>c. Different</td>
<td>oxytone - paroxytone</td>
</tr>
<tr>
<td>d. Different</td>
<td>paroxytone - oxytone</td>
</tr>
</tbody>
</table>

The target word pairs only differed in the placement of lexical stress. For example, they heard *pseudoword A* followed by *pseudoword B* and had to decide if the two words were identical. All of the pseudowords used in the task had two syllables and followed the phonotactic rules of either English or Spanish. The stimuli were recorded by a 25-year-old female simultaneous bilingual with a Sure SM10A head-mounted microphones and a Fostex DC-R302 digital recorder in a sound proof booth. Stimuli were randomly presented using PsychoPy3.
4.2.4 Data analysis

I report four primary statistical analyses to address the previously mentioned research questions. Two of the models analyzed response accuracy on the AX discrimination task and two different models analyzed sensitivity to lexical stress contrasts via $d'$ ($d$ prime). The analyses were conducted in R (R Core Team, 2019). The models were fit using *stan* (Stan Development Team, 2018) via the R package *brms* (Bürkner, 2017).

The response accuracy data were analyzed using Bayesian multilevel logistic regression in order to model the probability of a correct response as a function of language (Spanish, English), age, and proficiency. Given the binary nature of the participant responses, the model likelihood was Bernoulli distributed with a logit link function. Similarly, the $d'$ data were analyzed using Bayesian multilevel regression. In this case, the model likelihood was assumed to be Gaussian. The predictors were the same in both models. Language was sum coded ($\text{Spanish} = -1$, $\text{English} = 1$) and the continuous predictors were standardized by subtracting each value from the mean and dividing by the standard deviation. Group-level effects were also the same in both models; participants and items were grouping variables with the effect of language varying for participants.

Prior beliefs about the parameters being estimated were included using regularization, weakly informative priors (Gelman, Simpson, & Betancourt, 2017), and were model-specific. For the response accuracy model the priors were normally distributed.
and centered at 0 with a standard deviation of 3 for the model intercept and 0.5 for all population-level parameters. A Cauchy distribution was used for the SD prior, which was centered at 0 with a standard deviation of 0.2. The correlation prior for grouping effects was \( \text{lkj}(2) \). For the \( d' \) model the priors for the intercept and the population-parameters were distributed as normal with mean 1.5, scale 2.5 and mean 0, scale 1, respectively. The SD and \( \sigma \) priors used Cauchy priors centered at 0, scale 1, and, again, the correlation prior for grouping effects was \( \text{lkj}(2) \). Response accuracy models were fit with 2000 iterations (1000 warm-up), and the \( d' \) models were fit with 12000 iterations (11000 warm-up). Hamiltonian Monte-Carlo sampling was carried out with 4 chains distributed between 4 processing cores.

For statistical inferences I describe in detail the posterior probability distributions for each parameter of interest as well as the 95% highest density interval (HDI), and the maximum probability of effect (MPE). I established a region of practical equivalence (ROPE) of ±0.1 and report the percentage of the 95% HDI contained within the ROPE. Model goodness-of-fit was assessed using posterior predictive checks (see Appendix G).
4.3 Results

4.3.1 Response accuracy

The results of the response accuracy model showed that, on average, the children performed above chance, responding correctly approximately 65.25% of the time ($\beta = 0.63$, HDI = [0.34, 0.89], ROPE = 0, MPE = 1). The probability of responding correctly did not vary as a function of language ($\beta = -0.02$, HDI = [-0.26, 0.24], ROPE = 0.6, MPE = 0.57). Response accuracy did increase as a function of proficiency ($\beta = 0.88$, HDI = [0.64, 1.14], ROPE = 0, MPE = 1), with higher proficiency scores associated with higher response accuracy. Furthermore, age was negatively correlated with response accuracy ($\beta = -0.60$, HDI = [-0.88, -0.30], ROPE = 0, MPE = 1), suggesting that the higher proficiency children were not necessarily the oldest. There was no evidence suggesting higher order interactions. A summary of the model output is available in Table 4.2. Additionally, a forest plot illustrates the model summary in Figure 4.1. Figure 4.2 provides a triptych plot illustrating the effect of proficiency on response accuracy at the mean age $\pm 1$ SD.

Table 4.2: Model output of response accuracy as a function of age, proficiency, and language in children.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Median</th>
<th>95% CrI</th>
<th>MPE</th>
<th>%ROPE</th>
<th>Rhat</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.63</td>
<td>[0.336, 0.899]</td>
<td>1</td>
<td>0</td>
<td>1.004</td>
<td>1339.02</td>
</tr>
</tbody>
</table>
The adults displayed an overall response accuracy of approximately 93.46% ($\beta = 2.66$, HDI = [1.78, 3.55], ROPE = 0, MPE = 1). Similar to the children, the probability of responding correctly did not vary as a function of language ($\beta = 0.14$, HDI = [-0.39, 0.66], ROPE = 0.27, MPE = 0.69). Figure 4.3 plots the posterior probability distributions of the model. Table 4.3 includes the complete model summary.

Table 4.3: Numeric summary describing language effect on accuracy scores in adults.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Median</th>
<th>95% CrI</th>
<th>MPE</th>
<th>%ROPE</th>
<th>Rhat</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.66</td>
<td>[1.78, 3.55]</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1224.33</td>
</tr>
<tr>
<td>Lang. effect</td>
<td>0.14</td>
<td>[-0.39, 0.66]</td>
<td>0.69</td>
<td>0.27</td>
<td>1</td>
<td>1476.18</td>
</tr>
</tbody>
</table>

Figure 4.1: Forest plot of response accuracy as a function of language, age, and proficiency. Points represent posterior medians ±66% and 95% HDI. The shaded area around the null value (0) represents the region of practical equivalence (ROPE) of ±0.01.
Figure 4.2: Probability of a correct response as function of proficiency and language while holding age constant at -1, 0, and 1 standard deviations from the mean. The individual lines represent 150 draws from the posterior distribution. Population averages are represented by the white lines.
Figure 4.3: Three plots that illustrate language effect on accuracy scores in adults. Panel A represents the density of the posterior distribution of the model fitting language effect. Panel B represents the percentage of correct responses broken down by language. Panel C illustrates the language differences in accurate responses between English and Spanish.
4.3.2 Sensitivity to lexical stress

Similar to the response accuracy analysis, the model examining d’ showed that sensitivity to lexical stress varied as a function of age ($\beta = -1.46$, HDI = [-2.33, -0.62], ROPE = 0, MPE = 1) and proficiency ($\beta = 0.96$, HDI = [0.65, 1.28], ROPE = 0, MPE = 1). There was no evidence suggesting an effect of language ($\beta = -0.03$, HDI = [-0.18, 0.23], ROPE = 0.67, MPE = 0.61), nor were there any higher order interactions. A summary of the model output is available in Table 4.4, along with a forest plot in Figure 4.4. Figure 4.5 provides a triptych plot illustrating the effect of proficiency on sensitivity scores at the mean age ±1 SD.

Table 4.4: Numeric summary of the parameter estimates of the model examining d’ as a function of age, proficiency, and language.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Median</th>
<th>95% CrI</th>
<th>MPE</th>
<th>%ROPE</th>
<th>Rhat</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.068</td>
<td>[0.786, 1.339]</td>
<td>1</td>
<td>0</td>
<td>1.002</td>
<td>2352.631</td>
</tr>
<tr>
<td>z-Proficiency score</td>
<td>0.959</td>
<td>[0.645, 1.282]</td>
<td>1</td>
<td>0</td>
<td>1.001</td>
<td>2310.958</td>
</tr>
<tr>
<td>z-Age</td>
<td>-1.467</td>
<td>[-2.326, -0.624]</td>
<td>1</td>
<td>0</td>
<td>1.001</td>
<td>2344.282</td>
</tr>
<tr>
<td>Language effect</td>
<td>0.029</td>
<td>[-0.178, 0.233]</td>
<td>0.606</td>
<td>0.668</td>
<td>1</td>
<td>2185.568</td>
</tr>
<tr>
<td>z-Prof. x z-Age</td>
<td>-0.461</td>
<td>[-0.914, 0.011]</td>
<td>0.969</td>
<td>0.043</td>
<td>1</td>
<td>2494.174</td>
</tr>
<tr>
<td>z-Prof. x Lang.</td>
<td>-0.079</td>
<td>[-0.323, 0.145]</td>
<td>0.752</td>
<td>0.528</td>
<td>1</td>
<td>2284.673</td>
</tr>
<tr>
<td>z-Age x Lang.</td>
<td>0.255</td>
<td>[-0.38, 0.884]</td>
<td>0.781</td>
<td>0.189</td>
<td>1</td>
<td>2466.462</td>
</tr>
<tr>
<td>Proficiency x Age</td>
<td>0.176</td>
<td>[-0.154, 0.53]</td>
<td>0.854</td>
<td>0.299</td>
<td>1.001</td>
<td>2053.742</td>
</tr>
</tbody>
</table>
Figure 4.4: Summary of posterior model estimates. Points represent posterior means. The vertical discontinuous line represents the point null value (0) surrounded by a region of practical equivalence (ROPE) of ±0.1.
Figure 4.5: The triptych plot depicts effects of d’ score as a function of proficiency score while holding age constant at -1, 0, and 1 standard deviations from the mean. The gray dots represent individual participants, while the lines represent samples from the posterior distribution. The white lines represent population averages.
Given that (1) the adult group was extremely accurate in the AX discrimination task and (2) there was no evidence for a language effect in either group, I compared the adult average d’ score (mean = 2.45 ±0.15 SE) with the posterior distribution of the children. Specifically, I calculated the difference from the adult mean for all values in the posterior in order to test the hypothesis that the average difference between groups was greater than 0. In other words, I calculated a posterior distribution of plausible values for the difference between the children and the adults. The result is illustrated in Figure 4.6. From the plot, one can see that there is a clear difference between the average d’ score for the adults and the posteriors of the children in English and Spanish ($\beta = 1.25$, HDI = [1.04, 1.45], ROPE = 0, MPE = 1).

Figure 4.6: Difference in sensitivity to lexical stress between children and adults. The points represent the posterior medians ±66% and 95% HDI. The vertical dotted line and shaded grey rectangle represent the average d’ score ± SE of the mean for the adults.
4.4 Discussion

This study cross-sectionally examined how sensitivity to lexical stress develops across elementary school aged children in comparison with adult bilinguals. The participants completed an AX discrimination task with minimal pairs of pseudowords that only differed in lexical stress. The pseudowords were either English-like or Spanish-like and followed the phonotactic rules of their respective languages. The rest of the discussion will be framed around the 3 research questions that guided the study.

For the first research question of if child Spanish-English heritage speakers perceive stress contrasts in both of their languages, I analyzed response accuracy and sensitivity to lexical stress via d’ scores. There was no evidence of an effect of language in response accuracy or d’, which provides compelling evidence that children are accurate and sensitive to stress contrasts in both of their languages. In terms of accuracy, the children had comparatively low scores, though they did perform above chance, indicating that they are, indeed, able to perceive stress contrasts.

In addition to the accuracy results, a d’ score was calculated for every participant to measure sensitivity to stress. A d’ value of 3 is close to a perfect performance whereas a score of 0 suggests no sensitivity at all. As shown in Figure 4.6, the children’s average d’ score in both English and Spanish is likely between 1 and 1.5. This means they are performing above chance, but they are not highly sensitive. These results add to the already existing body of literature that support the idea that early exposure to
a language provides an advantage for perception abilities (Boomershine, 2013; Kim, 2015; Ronquest, 2013; Ronquest & Rao, 2018; Willis, 2005).

The second research question focuses on how proficiency and age mediate sensitivity to stress. We found evidence for a proficiency effect so that the higher proficiency the higher the sensitivity to lexical stress. This aligns with previous findings that show that the heritage speakers with less experience with Spanish (lower proficiency) or dominance in English show deviations from monolingual-like perception (Alvord, 2010b, 2010a; Henriksen, 2012; Rao & Ronquest, 2015).

Proficiency was also a positive predictor of accuracy responses, as children with higher proficiency were more accurate than those with a lower proficiency. This finding aligns well with previous research that extend L2 speech learning models to the acquisition of suprasegmentals (Trofimovich & Baker, 2006). L2 speech models, such as the PAM and SLM, assert that language experience (proficiency) and exposure, modulate L2 speech perception.

Finally, to answer the third research question that compares the children to the adults in terms of perception and sensitivity to stress, we found that adults are much more sensitive to stress contrasts than children are. For accuracy scores, the children are performing above chance, while the adults’ mean accuracy score is above 90%. To contextualize these findings, we can turn to Kim (2015) and Kim (2019). Taken together with the current study, the three experiments showed that all adult bilinguals and monolinguals perform similarly in terms of accuracy scores. In terms
of sensitivity, the present study revealed that adults are highly sensitive to minimal stress pairs. It is important to note that even at the highest proficiency level and highest age range, children are not as sensitive to minimal stress pairs as adults are. As a follow-up study, it would be prudent to explore further this development and track a larger range of ages. In doing so, we would be able to see at what point in their language acquisition do they reach the adult-like sensitivity to lexical stress found in the current study. This lower sensitivity in comparison with adults, however, could be due to the difficulty of the task. The use of pseudowords instead of real words should be taken into consideration when thinking about the children’s scores in comparison with the adults’. Pseudowords could have increased the difficulty of the experimental task because of the lack of familiarity with these lexical items. The novelty of the target words could have resulted in lower accuracy and sensitivity scores as compared to highly frequent real words in either English or Spanish.

4.5 Conclusion

The present chapter explored the development of perception of lexical stress in Spanish-English bilingual children in comparison to Spanish-English bilingual adults. Participants completed an AX discrimination task that aimed to see how sensitive they are to stress contrasts. First, the results of the accuracy scores showed that in comparison with the adults, the children had lower accuracy scores, but still performed above chance. In terms of sensitivity, the children are sensitive to minimal
stress pairs in pseudowords that were English-like and Spanish-like, but not as sen-
sitive as a bilingual adult comparison group. The children’s sensitivity to stress
contrasts supports the notion that early exposure to a language provides an advan-
tage for perception abilities. An effect of proficiency was found so that the higher
proficiency the higher the accuracy and the higher the sensitivity to lexical stress.
The effect of proficiency supports the extension of L2 speech learning models to the
acquisition of suprasegmentals because L2 models like the PAM and SLM, assert that
language experience (proficiency) modulate L2 speech perception. Finally, no effect
of language was found, as both adults and children were as accurate and as sensitive
to the stress contrasts in both English-like and Spanish-like pseudowords. The next,
and final experimental chapter, will explore the relationship between perception and
production and includes a teaching intervention task aimed at improving perception.
Chapter 5: Perceptual gains via instructional games: an examination of perception and production of lexical stress in heritage bilingual children

5.1 Introduction

Lack of formal education in a heritage language has often been used as an explanation for why there appears to be a difference in perception and production abilities between heritage speakers and their monolingual counterparts (Kupisch & Rothman, 2018). Heritage speakers are individuals who are “raised in a home where a non-English language is spoken, who speak or merely understand a heritage language and who are to some degree bilingual in English and in the heritage language” (Valdés, 2000). Throughout the lifespan of a heritage speaker, the use of English, the majority language, and the heritage language, Spanish in this context, can fluctuate with a tendency to use the majority language more often. More frequent use of the majority language has often resulted in a notable difference of linguistic skills in the heritage language when comparing heritage speakers to their parents or to peers raised in a monolingual setting Valdés (2000)

Literacy skills and the lack of a formal education in general in the heritage language have previously been associated with the production of read samples that are less natural (Ronquest & Rao, 2018). Heritage speakers acquire the heritage language naturalistically at home, and as a result of this naturalistic acquisition, heritage speakers tend to comprehend the language more than they can produce it because
they are often not required to respond back to parents or guardians in the heritage language Hurtado & Vega (2004). Unsurprisingly, a discrepancy between productive and receptive skills, with receptive skills being more comparable to a monolingual-like norm than productive skills, has been shown in numerous studies Kim (2019). For example, Kim (2019) found in Spanish heritage speakers, early exposure to the heritage language is beneficial for the perception of heritage language speech sounds, as they performed comparably to Spanish monolinguals when distinguishing stress minimal pairs. When producing the speech sounds, however, the heritage speakers behaved more similarly to the second language (L2) learners, indicating that early exposure alone is not enough to guarantee target-like production.

Though there is evidence of a discrepancy between productive and receptive skills in heritage speakers, there is an acute lack of research and research-based pedagogy in the area of phonological acquisition among Spanish heritage speakers. This exploratory study is an important step in understanding how perception of lexical stress changes across elementary school aged children and what learning strategies may contribute in that development and acquisition. This study aims to investigate the role that instructional games play in the development of the perception of lexical stress across elementary aged students, as it is an area with which heritage speakers, and L2 learners alike, struggle (Beaudrie, 2007; Carreira, 2002). Though lexical stress is a persistent difficulty for both groups of learners, with instruction an increase in sensitivity to lexical stress contrasts can be achieved (Romanelli, Menegotto, & Smyth,
I examine bilingual children’s perception of lexical stress contrasts in pseudowords and investigated how that perception changed after targeted instruction.

5.1.1 Lexical stress

English and Spanish both have lexical stress, which is the ability to form words with different meanings when the position of stress in alternated in a word (Hualde, 2005). This can be seen in the following example in Spanish:

a. hablo, /a.blo/, Eng. I speak

b. habló, /a.'blo/, Eng. He/she/it speaks

The difference between the example words, hablo and habló, can only be found in the position of the stressed syllable, which falls on the initial syllable in (a) making it a paroxytone and on the final syllable in (b) making it an oxytone. The placement of stress changes the meaning of the word completely. This type of contrast also exists in English, albeit to a lesser extent (e.g., “PROduce,” noun: harvested fruits and vegetables vs “proDUCE,” verb: to create something).

Although stress functions in a similar manner in both languages there are some key differences. Spanish is traditionally described as being a syllable-timed language in which every syllable is perceived as taking up roughly the same amount of time. Nonetheless, many studies have shown that lexical stress is primarily signaled by
duration Ortega-Llebaria & Prieto (2009). In contrast, English can be described as a stress-timed language, where stressed syllables are produced at approximately regular intervals and unstressed syllables reduce to fit that rhythm. Unlike Spanish, the suprasegmental cues that mark stressed syllables are a result of the primary cue that indicates the position of stress in an English word, i.e., vowel reduction (Ortega-Llebaria, Gu, & Fan, 2013).

With these differences in mind, it makes sense that previous research has found that monolingual speakers of English and monolingual speakers of Spanish process stress differently in their respective languages Cooper, Cutler, & Wales (2002). Evidence from cross-modal priming studies suggest that English and Spanish listeners weight suprasegmental cues in very different ways when processing auditory information Cooper, Cutler, & Wales (2002).

5.1.2 **Stress perception and language acquisition**

The manner in which the perception of speech sounds develops and is acquired in child heritage bilinguals has been largely under-researched. In contrast, the myriad of research done with infants shows that during the first year of life babies lose the ability to discern speech sounds from all languages and become more sensitive to the sounds relevant to their language. This ability applies to segmental Werker & Tees (1984) and suprasegmental levels of speech Skoruppa et al. (2013). L1 perception of stress has shown that suprasegmentals of a language, like stress patterns, are acquired early
on (Pons & Bosch, 2010). Factors such as syllable weight, morphological category, subregularities in the lexicon, and the influence of segmentally similar words play a role in the perception of word stress (Timothy Lee Face, 2000; Timothy L. Face, 2006; Pons & Bosch, 2010). English and Spanish speakers have been shown to process lexical stress differently in their respective L1s by using the suprasegmental cues available to them when processing auditory information in different ways (Cooper, Cutler, & Wales, 2002; Soto-Faraco, Sebastián-Gallés, & Cutler, 2001).

Research into L2 stress perception has attempted to account for the difficulties that language learners experience when perceiving stress in another language. One explanation is that the L1 prosodic system influences the perception of L2 phonology (Chrabaszcz, Winn, Lin, & Idsardi, 2014; Lee, Shin, & Garcia, 2019; Li & Grigos, 2018; Ortega-Llebaria, Gu, & Fan, 2013). Another factor that modulates L2 stress perception is early exposure. There is evidence to support the notion that the earlier the exposure to the second language, the better it is for more monolingual-like perception of stress (Dupoux, Peperkamp, & Sebastián-Gallés, 2010; Guion, Harada, & Clark, 2004).

Stress poses difficulties for heritage speakers, as identifying the stressed syllable in a word has been found to be the most common error for heritage speakers in writing evaluations (Beaudrie, 2007, 2017; Carreira, 2002). Kim (2015) and Kim (2019) explored this difficulty in how heritage speakers perceive stress through forced-choice identification tasks. The participants, heritage Spanish speakers who were dominant
in English, were tasked with choosing the subject of the sentence, either *yo* (I) or *él/ella* (he/she). The results of the two studies show that heritage speakers were performing at ceiling, just like the monolinguals in the study. Overall, the conclusions drawn from the two aforementioned studies support the notion that early exposure to a language likely aides in monolingual-like perception of stress.

### 5.1.3 Production of lexical stress

As in L2 perception, L2 production of stress is also influenced by the L1 and early exposure. Influence from the L1 prosodic system has been shown to lead to less target-like stress pattern production in L2 English (Li & Grigos, 2018; Saha & Mandal, 2016). This includes, for example, differences in duration, intensity, pitch, and vowel quality. Additionally, target-like production skills have been shown in some studies, but not in others Flege & Bohn (1989a). Flege & Bohn (1989a) explored how native speakers of Spanish produced vowel reduction and stress placement in English. The results of the study showed that producing stress placement caused less problems than vowel reduction for L2 learners, suggesting that English-like stress placement is acquired earlier than vowel reduction for L1 Spanish speakers. In regard to age, L2 production also seems to be modulated by early exposure. Guion, Harada, & Clark (2004) investigated early and late Spanish-English bilinguals’ acquisition of English word stress. The results of the study showed non target-like production of stress in non-words, leading to the conclusion that early exposure might not be as big of an
indicator of target-like production as it is for perception of stress.

In the same vein, heritage speakers, who have early exposure to their heritage language, also exhibit difficulties in regard to production of stress (Beaudrie, 2012, 2017; Carreira, 2002; Kim, 2015, 2019; Rao & Kuder, 2016). Kim (2015) and Kim (2019) compared production abilities of lexical stress in heritage speakers, L2 learners, and monolingual Spanish speakers. In both studies, participants were tasked with reading sentences out loud that were shown to them on a computer screen. Results from Kim (2015) revealed that in nearly half of the cases, heritage speakers and L2 learners produced the unstressed vowels longer than the stressed vowels in words with paroxytonic stress. Adding to this deviation from monolingual-like production of stress, Kim (2019) found that heritage speakers largely deviated from monolinguals and performed more similarly to the L2 learners when producing Spanish stress minimal pairs. These results contribute more evidence suggesting early exposure does not necessarily equate to advantages in terms of production abilities. The two studies together support for the notion that there is often a discrepancy between production and perception abilities in heritage speakers. Specifically, the ability to perceive speech in the heritage language appears to exceed the ability to produce it.

5.1.4 Perception and production interface

The nature of the relationship between the perception and production is debated despite being heavily researched (Akahane-Yamada, Tohkura, Bradlow, & Pisoni,
110

1996; Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Flege, 2003; Goto, 1971; Leather, 1997; Matthews, 1997; Sheldon & Strange, 1982; Strange, 1995; Wang, 2002; Zampini, 1998). The debate centers around the question of if the relationship is a causal one, and if it is, which modality causes the other to change and develop. It is also up for debate as to whether or not a relationship even exists between the two modalities. Models of speech perception, such as Motor Theory, Direct Realist Theory, and General Auditory and Learning Approaches, all propose a fairly simple relationship between perception and production. This seemingly simple relationship, however, is not as straightforward when taking L2 language learners into consideration.

To begin with the argument that the relationship is a causal one in L2 acquisition, the Speech Learning Model (SLM) maintains that L2 phonetic segments are produced only as accurately as they are perceived (Flege, 2003). Evidence for this claim can be found in studies that have shown that perceptual shifts often lead to gains in production (Akahane-Yamada, Tohkura, Bradlow, & Pisoni, 1996; Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997). Conversely, there is evidence for the claim that the causal relationship goes in the opposite direction, that production drives perception (Goto, 1971; Sheldon & Strange, 1982). A semi-longitudinal study that investigated the Spanish stop contrast /p/ - /b/ revealed that production changes occurred before perceptual changes did (Zampini, 1998). This study, along with other studies that explore how explicit instruction for L2 production can drive changes
in perception is direct evidence against the predictions of the SLM (Leather, 1997; Matthews, 1997).

The relationship between the two, however, is not as clear-cut as the previously mentioned studies seem to suggest. Improvements in perception do not always lead to improvements in production Strange (1995). Wang (2002) studied the perception and production of American English vowel contrasts in adult L1 Mandarin L2 English learners. This research showed that while the participants improved in their perception of contrasts, their production did not improve. Additionally, Strange (1995) claims that perceptual changes generally occur more slowly over time when input and use are factored in.

In sum, the nature of the relationship between the modalities of perception and production is a complex one that remains a subject of debate. There is evidence for changes in one driving changes in the other, making it a causal relationship. There is also evidence for a not so defined relationship between the two, where changes in one do not impact the other. The current study adopts the SLM’s position that perception drives production and implements a teaching intervention aimed at improving perception skills.

5.1.5 Bilingual language education

Due to the fact that discerning speech sounds, like lexical stress, poses some difficulties for both L2 learners and heritage speakers, it is prudent to see how teaching
interventions and instruction could improve bilingual children’s perception of lexical stress. There is a lack of research-based practices in relation to phonology in the elementary-aged heritage language classroom. In general, pedagogical interventions have been shown to increase sensitivity to perception of lexical stress in adult L2 and heritage speakers (Beaudrie, 2017; Bullock & Lord, 2003; Timothy L. Face, 2005; Gutiérrez-Palma, Naranjo, Justicia-Galiano, & Fernández, 2019; McAndrews, 2019; Romanelli, Menegotto, & Smyth, 2015). To directly measure the impact of level of instruction on perception, Romanelli, Menegotto, & Smyth (2015) and Beaudrie (2017) both used a pre/post test design to examine perception of lexical stress. Romanelli, Menegotto, & Smyth (2015) examined the perception of Spanish vowels and lexical stress in beginner L1 English learners of Spanish in an immersion setting. The pretest results showed that perception of stress contrasts is a difficult challenge for the learners, while the post-test results showed that English speakers learned to perceive penultimate stress like natives but continued to identify final stress inaccurately. For heritage speakers, Beaudrie (2017) evaluated a step-by-step instructional approach that focused on stress identification through three computer-based training sessions. The heritage speakers in this study completed a pre-test and a post-test after six weeks of instruction. The results of the study showed significant progress towards the mastery of stress identification as a result of the instruction. Taken together, the studies show that both heritage speakers and L2 learners benefit from explicit instruction and a teaching intervention for Spanish stress.
Importantly, instruction does not have to be lengthy for an increase in sensitivity to be found. McAndrews (2019)'s comprehensive review study on the effects of instruction on suprasegmental listening skills aimed to determine how effective instruction is on improving these skills. The results of the meta-analysis indicated that instruction has a large and positive impact on the learner’s development of phonological categories for suprasegmental features in the target language. Additionally, effects were found with as little as one hour of instruction.

5.1.6 Current Study

The present study aims to conceptually replicate Kim (2019) and explore the development of the perception and production of lexical stress in Spanish and English in bilingual children. Lexical stress was chosen because it has been shown to be persistently difficult for L2 and heritage speakers to perceive and produce in a monolingual-like way (Beaudrie, 2012, 2017; Carreira, 2002; Rao & Kuder, 2016). Previous research into lexical stress has focused on adult L2 learners and heritage speakers (Beaudrie, 2017; Guion, Harada, & Clark, 2004; Kim, 2015, 2019; Li & Grigos, 2018; Romanelli, Menegotto, & Smyth, 2015; Saha & Mandal, 2016; Zhang, Nissen, & Francis, 2008). This cross-sectional study of the development lexical stress in heritage speakers allows for the exploration of how both languages influence each other during acquisition across the elementary school ages. This study also has a second aim of exploring how pedagogical practices aid in the perception of lexical stress across elementary-aged
students. Though lexical stress is a persistent difficulty for both heritage speakers and L2 learners, previous research suggests that with instruction an increase in sensitivity to lexical stress contrasts can be achieved (Beaudrie, 2017; McAndrews, 2019; Romanelli, Menegotto, & Smyth, 2015). I examine bilingual children’s perception of lexical stress contrasts in pseudowords and investigate how that perception changed after an hour of instruction.

The following questions guide the study:

1. Can lexical stress perception be improved using pedagogical interventions?
2. Does improvement depend on stress pattern and language?
3. Are there individual differences in learning based on age or proficiency?
4. Is there a discrepancy between perception and production abilities?

Following previous research, I hypothesize that stress perception will be improved as a result of the instructional games aimed at developing perception skills. As the instruction will be given in Spanish, I expect the improvement to depend on language but not on stress pattern. Additionally, I hypothesize that individual differences in learning will be based on both age and proficiency. Finally, as with previous studies, I expect the bilinguals to have better receptive than productive skills in comparison to an adult bilingual group.
5.2 Methods

A total of 39 participants completed five tasks over the course of several days in the following order: (1) linguistic background questionnaire, (2) proficiency task, (3) a AX discrimination task, (4) a production task, and (5) a teaching intervention.

5.2.1 Participants

There were a total of 39 participants, 32 Spanish heritage speaker children (n = 32, ages: 6.6 - 11.7) and a comparison group of adult bilinguals (n = 7). All of the children that participated in this study were recruited from dual language immersion schools servicing students kindergarten through 8th grade in New Jersey. The heritage speakers were born and raised in the US, mostly in the central region of New Jersey. All heritage speakers were consecutive bilinguals who had learned and acquired both English and Spanish before the age of five. All participants provided information regarding the contexts, amount per week, and people with whom they speak each language. Participants were measured on their phonological abilities through a standardized test. Additionally, they completed the Bilingual Language Profile (BLP) in order to assess language history, proficiency, and use. None of the participants has working knowledge of any language other than Spanish or English.
5.2.2 Screening Tasks

The following screening tasks were implemented in order to ensure that the participants were all evenly matched in terms of input, exposure, and proficiency. First, the BLP elicited background language information relevant to the research questions (e.g., age of onset of acquisition of Spanish, context of language use, etc.). Second, the proficiency task, is a standardized test of phonological knowledge/ability currently being used by the school that the students are attending. The results of this test are used to measure participants’ proficiency in Spanish.

5.2.2.1 Background questionnaire

The background questionnaire—the BLP—was completed by the participants’ parents, along with the informed consent form. The BLP gathered information about the language history, proficiency, and use in the two languages through self-reports and will measure of relative language dominance. The questionnaire probes participants regarding their experience and exposure to language, including age of exposure to Spanish and English, number of years in formal education in each language, location, and length of time living abroad. Subsequently, a self-report Likert scale assessment of their Spanish and English abilities in reading, writing, speaking, and listening was filled out by the children. There was also questions about with who, in what context, and when they speak each language, with the goal of ensuring homogeneity among
the participants and controlling for possible confounds.

5.2.2.2 Proficiency Task

The students all completed an assessment of their phonological abilities that I used as a proxy for proficiency. The assessment is one that the school has implemented and uses as a means of tailoring instruction to meet the needs to the students. The test used is the Test of Phonological Awareness (TOPA-2+), which measured phonological awareness. The TOPA-2+ has two versions, a Kindergarten version and an Early Elementary version, that measure young children’s ability to (a) isolate individual phonemes in spoken words and (b) understand the relationships between letters and phonemes in English. The test consists of numerous parts, however, for the purposes of the present study I focus on two: a final sound identification task and a Letter-Sound test where students spell simple pseudowords. The school district adopted this test in order to keep track of student progress.

5.2.3 Experimental tasks

Two experimental tasks are presented. First, an AX discrimination task was implemented in a pre/post test design in order to assess the participants abilities to discern differences between oxytonic and paroxytonic stress. Second, a delayed-repetition
production task (DRT) examined how the participants produced lexical stress.\(^1\) Crucially, a teaching intervention took place between pre and post sessions of the AX discrimination task, which allowed participants to practice, develop, and improve on their perception of stress.

### 5.2.3.1 Perception: AX discrimination

An AX discrimination task was administered to assess the participants' sensitivity to lexical stress contrasts. The experiment was programmed through PsychoPy3 software and was hosted online (Peirce et al., 2019). The participants were told that they were going to hear a series of rare words. Their task was to decide if the words that they heard were the same or different. They first saw a red cross in the middle of the screen for 150 ms to draw their attention and to signify the start of a new trial. Next, they heard the two target words in a sequence. After the presentation of the oral stimuli, the participants saw the words “same” and “different” on the screen. The participants were instructed to keep their fingers on keys “1” and “0” on their keyboards to record their responses. The pairs of words that they heard only differed in the placement of lexical stress. For example, they heard *pseudoword A* followed by *pseudoword B* and had to decide if the two words were identical. All pseudowords have two syllables and follow the phonotactic rules of either English or Spanish. The

\(^1\)A production post test was not included in this study due to conflicts in scheduling and meeting spaces that arose due to the COVID-19 pandemic.
stimuli were presented in one of either four conditions:

Table 5.1: Example trial types used in the AX discrimination task.

<table>
<thead>
<tr>
<th>Status</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Same</td>
<td>oxytone - oxytone</td>
</tr>
<tr>
<td>b. Same</td>
<td>paroxytone - paroxytone</td>
</tr>
<tr>
<td>c. Different</td>
<td>oxytone - paroxytone</td>
</tr>
<tr>
<td>d. Different</td>
<td>paroxytone - oxytone</td>
</tr>
</tbody>
</table>

The stimuli were recorded by a 25-year-old female simultaneous bilingual with a Sure SM10A head-mounted microphones and a Fostex DC-R302 digital recorder in a sound proof booth, and presented randomly using PsychoPy3.

5.2.3.2 Delayed repetition task

The DRT was conducted both in English and Spanish. Here, participants were presented with auditory and visual stimuli. They simultaneously heard a phrase being read while it was presented on a screen. Once the sentence finished playing, they were required to repeat back what they heard and saw. The sentence stayed on the screen so that they could refer back to it as needed. In total, there were 32 sentences that contained 3-6 words each with either a past or present tense verb. Target words never came first or last in a sentence. The following are samples of the task in both
English and Spanish.

Table 5.2: Examples of target items in the delayed repetition task.

<table>
<thead>
<tr>
<th>Example item</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Hoy trabajo contigo. (Today I work with you.)</td>
</tr>
<tr>
<td>b. El hombre trabajó ayer. (The man worked yesterday)</td>
</tr>
<tr>
<td>c. There is a present here.</td>
</tr>
<tr>
<td>d. I have to present today.</td>
</tr>
</tbody>
</table>

Participants were recorded with a Fostex DC-R302 digital recorder and the sound files were later segmented and labeled using PRAAT.

5.2.3.3 Teaching intervention

This project aimed to see how a teaching intervention could help the participants improve their perceptual abilities in the previously described AX discrimination task. For 20 min a day over the course of a week, the students played one of the following games in Spanish:

1. Stress Bingo: This game had the students to categorize words based on where the tonic syllable falls in a word. This game forced the students to pay attention to where the stressed syllable is in a word. This game was played through Zoom because schools were shut down during data collection due to COVID-19. For
this version of Bingo, there were only two columns that corresponded to either paroxytone or oxytone stress pattern. They were labeled as oO or Oo. The students were tasked with classifying the words that the researcher called out in either column. For example, the researcher would call out “mamá” (Eng. *mom*), and the students would have to put that word in the oO column.

2. Flyswatter Game: This is a typical game in the foreign language classroom. This game was also played through Zoom with more than one student in the meeting. If there were more than two players, they were divided into two teams, but only two students competed at a time. The researcher said a word and the two students who were competing had to answer as quickly as possible with the answer that matches which syllable is the stress syllable in the word. For example, the research said “cantó” (Eng. *he/she/it sang*), and the students had to yell out “smallBIG.”

3. Silent Ball: This was the only game played in person before data collection was interrupted due to the COVID-19 pandemic. As with the aforementioned games, the goal of this game was to have the participants practice identification of stressed vs non-stressed syllables. Every student was a “weak/small syllable” or a “strong/big syllable.” The researcher said a word out loud and the student with the ball had to say the first syllable in that word and throw the ball to a student that matches the description of the syllable. For example, the researcher said “habló” (Eng. *he/she/it spoke*) and the student said “can” and throws the
ball to a weak/small syllable student. The student that gets the ball then says “tó” and throws it to a “strong/big” syllable student. If they throw it to the wrong syllable type person, they’re out.

After 5 days of playing games for short sessions (totaling 1-2 hours of instruction), the students completed the second implementation of the AX discrimination task.

5.2.3.4 Power Analysis

An a priori power analysis was performed for sample size estimation. The complete code along with a detailed explanation are available in Appendix I. Based on the literature and similar studies, a small/moderate effect size of 0.3 was chosen. The alpha value is the standard 0.05. The results of the analysis suggest that the present study would need a sample size of at least 38 participants to reach 0.8 power. Figure 5.1 plots the results of the power analysis.

5.2.4 Data analysis

The results section includes two subsections, one for the perception study and another for the production study. At the beginning of each section I provide critical details necessary to understand each model. Here I describe the basic structure that was applied to all statistical analyses. The data were analyzed using Bayesian multilevel regression models. The models examined how perception and production metrics
Figure 5.1: Power analysis for pre/post test discrimination task. Change in d’ scores as a function of experimental session (simulated data). The points represent the mean scores. The different shades of blue correspond to the percentage of participants in that range of d’ scores. For example, the darkest shade of blue represents 50% of the participants, where as the lightest shade of blue represents 95% of the participants.
varied as a function of a series of predictors (described below). In all cases I fit the maximal model justified by the structure of the data (Barr, Levy, Scheepers, & Tily, 2013). The analyses were conducted in R (R Core Team, 2019) and models were fit using `stan` (Stan Development Team, 2018) via the R package `brms` (Bürkner, 2017). Prior beliefs about the parameters being estimated were included using regularizing, weakly informative priors (Gelman, Simpson, & Betancourt, 2017). Hamiltonian Monte-Carlo sampling was carried out with 4 chains distributed between 4 processing cores. As a result, all posterior distributions estimated consisted of 4,000 samples that were compatible with the data and our prior beliefs. For statistical inferences I describe in detail the posterior probability distributions for each parameter of interest, as well as the 95% highest density interval (HDI), and the maximum probability of effect (MPE). I established a region of practical equivalence (ROPE) of $\pm 0.1$ and report the percentage of the 95% HDI contained within the ROPE. The goodness-of-fit of each of the models was checked using posterior predictive checks (see Appendix H).
5.3 Results

5.3.1 Effects of pedagogical interventions on perception of lexical stress

5.3.1.1 Accuracy

The first series of analyses modeled response accuracy using Bayesian multilevel logistic regression. The model analyzed the probability of a correct response as a function of language (Spanish, English), age, and proficiency. Given the binary nature of the participant responses, the model likelihood was Bernoulli distributed with a logit link function. Language was sum coded (Spanish = -1, English = 1) and the continuous predictors were standardized by subtracting each value from the mean and dividing by the standard deviation.

Overall the children displayed a large amount of variability in their response accuracy. The average score was approximately 80.38%, though credible interval around this estimate was particularly wide. We can say with 95% certainty that the median value falls between a lower bound of 39.65% and an upper bound of 96.19%. The teaching intervention improved accuracy (time: $\beta = 0.39$, HDI = [0.03, 0.68], ROPE = 0.03, MPE = 0.98), though a small portion of the plausible values for the effect fell within our pre-established ROPE (approximately 3% of the HDI). There was no evidence for an effect of language nor was there a intervention $\times$ language interaction. Figure 5.2 illustrates the pre/post task effect on accuracy scores and Table 5.3
summarizes the model output.

Figure 5.2: Response accuracy pre- and post-test. Two panels represent accuracy on English- or Spanish-like pseudowords. The points represent the posterior medians ±66% and 95% HDI.

Table 5.3: Numeric summary of the posterior distributions of the model describing response accuracy.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Median</th>
<th>95% CrI</th>
<th>MPE</th>
<th>%ROPE</th>
<th>Rhat</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.41</td>
<td>[-0.42, 3.23]</td>
<td>0.94</td>
<td>0.03</td>
<td>1</td>
<td>2222.13</td>
</tr>
<tr>
<td>Intervention</td>
<td>0.39</td>
<td>[0.02, 0.68]</td>
<td>0.98</td>
<td>0.03</td>
<td>1</td>
<td>2783.49</td>
</tr>
<tr>
<td>Language</td>
<td>0.02</td>
<td>[-0.41, 0.42]</td>
<td>0.53</td>
<td>0.41</td>
<td>1</td>
<td>2349.57</td>
</tr>
<tr>
<td>Intervention x Language</td>
<td>-0.02</td>
<td>[-0.14, 0.09]</td>
<td>0.63</td>
<td>0.93</td>
<td>1</td>
<td>6962.37</td>
</tr>
</tbody>
</table>

Looking deeper into the improvement made across stress patterns and language,
one observes that the order in which the stimuli were presented, (oxytone-paroxytone, paroxytone-oxytone, oxytone-oxytone, and paroxytone-paroxytone) affected response accuracy (see Figure 5.3). In both English and Spanish, participants had a harder time accurately responding when they were presented with diverging stress patterns and when oxytone words came first. Additionally, improvements were made for correctly responding “different” to different stress patterns presented in the same trial.

Figure 5.3: Response accuracy as a function of stress patterns. The two “different” categories represent trials that presented the participants with either a minimal pair of oxytone-paroxytone words or paroxytone-oxytone words. The points represent posterior medians ±66% and 95% HDI.
5.3.1.2 Effect of teaching interventions on sensitivity to lexical stress

The next analysis fit the same model as described in the previous section, with the exception that the outcome variable was d’ . In this case, the model likelihood was assumed to be Gaussian. Figure 5.4 shows d’ scores before and after the teaching intervention in both languages.

![Figure 5.4: Summary of the change in d’ score between the pre- and post-tests. The plot depicts change in sensitivity to English- and Spanish-like pseudowords. The lines represent individual participants and their change in d’ score. The points represent posterior medians ±66% and 95% HDI.](image)

Similar to the accuracy results, the intervention effect (pre vs. post) accounted for increased sensitivity to lexical stress ($\beta = 0.31$, HDI = [0.20, 0.41], ROPE = 0, MPE = 1); however, the effect was substantially less uncertainty around the estimate. Language and the interaction between language and the intervention did not con-
tribute to changes in sensitivity. The numerical summary of the output of the model can be seen in Table 5.4.

Table 5.4: Numeric summary of the posterior distributions that describe change in sensitivity to lexical stress.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Median</th>
<th>95% CrI</th>
<th>MPE</th>
<th>%ROPE</th>
<th>Rhat</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.79</td>
<td>[1.65, 1.94]</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1871.9</td>
</tr>
<tr>
<td>Intervention</td>
<td>0.31</td>
<td>[0.2, 0.41]</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3764.6</td>
</tr>
<tr>
<td>Language</td>
<td>0.03</td>
<td>[-0.06, 0.12]</td>
<td>0.73</td>
<td>0.96</td>
<td>1</td>
<td>5757.56</td>
</tr>
<tr>
<td>Intervention x Language</td>
<td>-0.01</td>
<td>[-0.1, 0.08]</td>
<td>0.64</td>
<td>1</td>
<td>1</td>
<td>6498.31</td>
</tr>
</tbody>
</table>

5.3.1.3 Effects of proficiency and age

The final model included the predictors age and proficiency to see if they accounted for changes in sensitivity to lexical stress. To this end, pre-test $d'$ scores were subtracted from post-test scores to calculate change scores (henceforth $\Delta d'$). $\Delta d'$ was the outcome variable for the final model and the continuous predictors were standardized by subtracting each value from the mean and dividing by the standard deviation. Figure 5.5 illustrates $\Delta d'$ as a function of standardized age and proficiency scores.

Given the model, the data, and our prior beliefs, we found moderate evidence suggesting change in sensitivity to lexical stress was moderated by age ($\beta = 0.45$, HDI = [0.03, 0.86], ROPE = 0.02, MPE = 0.98). We can say with 98% certainty
Figure 5.5: $\Delta d'$ as a function of age while holding proficiency constant at -1, 0, and 1 standard deviations from the mean. The points in the plot represent individual participants whereas lines represent 150 draws from the posterior distribution.
that the effect is positive and approximately 2% of the HDI fell within the ROPE. Language and proficiency score were not found to be predictors of change in d’, though there was also moderate evidence for the interaction between age and proficiency score ($\beta = 0.34$, HDI = [0.07, 0.59], ROPE = 0.02, MPE = 0.99). In this case, we can say with 99% certainty that the interaction effect was positive and approximately 2% of the HDI fell within the ROPE. The numerical summary of the output of the model is provided in 5.5.

Table 5.5: Numeric summary of the posterior distributions describing the change in sensitivity to lexical stress. The table includes all possible interactions between age, proficiency score, and language.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Median</th>
<th>95% CrI</th>
<th>MPE</th>
<th>%ROPE</th>
<th>Rhat</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.32</td>
<td>[0.05, 0.61]</td>
<td>0.99</td>
<td>0.04</td>
<td>1</td>
<td>2810.01</td>
</tr>
<tr>
<td>Language</td>
<td>-0.01</td>
<td>[-0.28, 0.29]</td>
<td>0.52</td>
<td>0.55</td>
<td>1</td>
<td>2210.66</td>
</tr>
<tr>
<td>z-Age</td>
<td>0.45</td>
<td>[0.03, 0.86]</td>
<td>0.98</td>
<td>0.02</td>
<td>1</td>
<td>1908.3</td>
</tr>
<tr>
<td>z-Prof.</td>
<td>-0.37</td>
<td>[-0.81, 0.08]</td>
<td>0.95</td>
<td>0.09</td>
<td>1</td>
<td>1828.64</td>
</tr>
<tr>
<td>Language x z-Age</td>
<td>-0.04</td>
<td>[-0.48, 0.35]</td>
<td>0.59</td>
<td>0.39</td>
<td>1</td>
<td>2080.77</td>
</tr>
<tr>
<td>Language x z-Prof.</td>
<td>-0.02</td>
<td>[-0.48, 0.41]</td>
<td>0.54</td>
<td>0.37</td>
<td>1</td>
<td>1948.7</td>
</tr>
<tr>
<td>z-Age x z-Prof.</td>
<td>0.34</td>
<td>[0.07, 0.59]</td>
<td>0.99</td>
<td>0.02</td>
<td>1</td>
<td>2511.76</td>
</tr>
<tr>
<td>Language x z-Age x z-Prof.</td>
<td>-0.03</td>
<td>[-0.29, 0.23]</td>
<td>0.58</td>
<td>0.57</td>
<td>1</td>
<td>2081.1</td>
</tr>
</tbody>
</table>
5.3.2 Production of lexical stress

The first of the series of production analyses modeled the standardized values of the V1-V2 differences of each production metric (duration, F0, intensity) as a function of lexical stress (paroxytone, oxytone) and language (Spanish, English). The metric, level of stress, and level of language were coded into an indicator variable and the omnibus model analyzed $\Delta V1 - V2$ for each indicator. The model did not include a population intercept, but there were grouping factors for participants and word items. The production results are summarized in Figure 5.6 which shows the posterior distributions of each parameter from the model. Table 5.6 provides a numeric summary of the posterior distribution.

Duration appears to be the primary cue that is used to produce V1-V2 differences. For English, paroxytonic words are clearly marked by duration differences ($\beta = 0.87$, HDI = [0.52, 1.22], ROPE = 0, MPE = 1), but oxytonic words to lesser extent ($\beta = -0.3$, HDI = [-0.6, 0.01], ROPE = 0.08, MPE = 0.97). In the case of the latter, there is a 97% chance that the effect is positive and approximately 8% of the HDI fell within the ROPE. For Spanish, there were moderate effects for both stress patterns, though is was stronger for oxytones ($\beta = -0.57$, HDI = [-1.14, -0.06], ROPE = 0.01, MPE = 0.98) than for paroxytones ($\beta = 0.33$, HDI = [-0.10, 0.75], ROPE = 0.11, MPE = 0.94). Intensity is used as a secondary cue in a manner that is language specific, that is, there was an effect paroxytones in English ($\beta = 0.70$, HDI = [0.25, 1.1], ROPE =
0, MPE = 0.99) and oxytones in Spanish (β = -0.50, HDI = [-0.86, -0.16], ROPE = 0, MPE = 0.99). The children did not use F0 to mark stress contrasts.

![Figure 5.6: ∆V1−V2 as a function of production metrics, stress, and language. Points represent posterior medians ±66% and 95% HDI. The vertical discontinuous line represents the point of null value (0) surrounded by a region of practical equivalence (ROPE) of ±0.1.](image)

<table>
<thead>
<tr>
<th>Language</th>
<th>Metric</th>
<th>Stress</th>
<th>Median (95% CrI)</th>
<th>MPE</th>
<th>%ROPE</th>
<th>Rhat</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Dur.</td>
<td>Ox.</td>
<td>-0.3 [-0.6, 0.01]</td>
<td>0.97</td>
<td>0.08</td>
<td>1</td>
<td>3742.4</td>
</tr>
<tr>
<td></td>
<td>Dur.</td>
<td>Parox.</td>
<td>0.87 [0.52, 1.22]</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3414.64</td>
</tr>
<tr>
<td></td>
<td>F0</td>
<td>Ox.</td>
<td>-0.07 [-0.4, 0.29]</td>
<td>0.66</td>
<td>0.42</td>
<td>1</td>
<td>1863.32</td>
</tr>
</tbody>
</table>
The next analysis fit the same model as described in the previous section with the addition of an age and age $\times$ indicator predictors. Age was standardized and included in the model as a continuous predictor. Figure 5.7 summarizes the posterior distributions of each parameter from the aforementioned model. The results suggest that the use of intensity, F0, and duration remains relatively constant across the ages that were tested in this experiment. Table 5.7 provides a numeric summary of the model output.

Table 5.7: Model summary of the posterior distribution modeling V1-V2 difference as a function language, stress, and age for each production metric.

<table>
<thead>
<tr>
<th>Language</th>
<th>Metric</th>
<th>Stress</th>
<th>Median</th>
<th>95% CrI</th>
<th>MPE</th>
<th>%ROPE</th>
<th>Rhat</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Dur. x z-Age</td>
<td>Ox.</td>
<td>-0.03</td>
<td>[-0.33, 0.27]</td>
<td>0.59</td>
<td>0.51</td>
<td>1</td>
<td>1578.18</td>
</tr>
<tr>
<td></td>
<td>Dur. x z-Age</td>
<td>Parox.</td>
<td>0.04</td>
<td>[-0.44, 0.47]</td>
<td>0.57</td>
<td>0.35</td>
<td>1</td>
<td>2242.8</td>
</tr>
</tbody>
</table>
5.4 Discussion

The present study conceptually replicated Kim (2019) and investigated the perception and production of lexical stress in Spanish and English bilingual children. Additionally, we measured how a teaching intervention can improve on perception skills. The participants completed four tasks: an AX discrimination task with minimal pairs of pseudowords, a DRT production task, a week-long teaching intervention, and finally repeated the initial AX discrimination task. The rest of the discussion will be framed around the research questions that guided the study.

The first question asked if lexical stress perception can be improved using pedagogical interventions. Our findings, specifically the results of the change in accuracy and
Figure 5.7: V1-V2 differences as a function of production metrics, stress, language, and age for each production metric.
d’ between the pre- and post-AX discrimination tasks, lead us to believe that lexical stress perception can be improved with a teaching intervention. The average change in d’ and accuracy scores was positive, showing that after a week of instruction, students were more sensitive to lexical stress and more accurate with their responses. This supports previous research on the impact of heritage language and phonology instruction in the classroom (Beaudrie, 2017; Romanelli, Menegotto, & Smyth, 2015), and research in general on phonological instruction improving speech perception (McAndrews, 2019). Such gains in perception support the push for instruction centered on improving phonological development in the language classroom, and especially in the heritage language classroom. The type of instruction is important to note, as each of the instructional games had a focus on form. Due to the fact that phonologically contrastive words have different meanings, giving instruction with a focus on meaning instead of form would also be another option for a pedagogical intervention for future research. A comparison between instruction with a focus on form and a focus on meaning would give insight into which type of instruction is better suited for increasing sensitivity to lexical stress, especially in the context of the current study which used pseudowords instead of real words. Additionally, because this was done via Zoom, it would be interesting to see how in person instruction would be different, if there is a difference, in the amount of improvement made. Furthermore, the post-test was administered after a week-long instructional intervention, but there was no delayed post-test. An addition of a delayed post-test would allow for the measure of
long-term impact of the teaching intervention.

Turning to the second research question of does improvement depend on stress pattern and language, we can look specifically at 5.3. It is evident that across both languages the combination of hearing an oxytone word first followed by a paroxytone word causes more difficulty for participants than the combination of a paroxytone word followed by an oxytone word. This difficulty could be attributed to the fact that paroxytone stress is the more common stress pattern in both English and Spanish. The participants could have been anticipating paroxytone words more often and when presented with the combination of an oxytone followed by a paroxytone, could have fixated on the paroxytone because it is the expected word stress and answered incorrectly. In general, improvement is possible across all languages and stress patterns. No effect of language is interesting to note in the context of the language used for the instructional games. All games were played in Spanish, but improvement was made in both English and Spanish. The improvement of both languages despite instruction in just Spanish provides evidence for the influence of the L1 on L2 perception.

The third research question asked if there are individual differences in learning based on age or proficiency. The results show that age and proficiency are both predictors of sensitivity to stress. Age and proficiency being a predictor of sensitivity aligns with previous studies (Kim (2019); Dupoux, Peperkamp, & Sebastián-Gallés (2010); Guion, Harada, & Clark (2004)). Furthermore, not all students improved on their scores between tests or improved by the same amount. This data provides
Finally, the last research question asked if there was a discrepancy between perception and production abilities. The production data show that the children are not using any specific cues to denote lexical stress, rather there are context specific uses for each cue. Duration is a primary cue, but moreso for paroxytones and is used to a lesser extent for oxytones. F0 is not used in either stress type or either language, and intensity is a secondary cue for paroxytones in English and oxytones in Spanish.

To compare the production data with the perception data, the response accuracy is low for the AX task. Given that this subset of participants are not making use of all of the cues available to them in production and are not highly accurate in their perception of stress, I can hypothesize that they have not completely developed the perception and production strategies necessary to discriminate and produce minimal stress pairs. It would be interesting to track this development further and see what changes are made to their production as they go through middle and high school. In doing so, we would be able to see at what point, if at all, they make the change in order to use all available cues at their disposal.

Furthermore, a post-test production task is the missing piece here to explore the relationship between perception and production. Due to conflicts in scheduling and audio recordings via Zoom that arose because of the COVID-19, a post-test in production was not possible. A post-teaching production task would help answer the question as to whether or not changes in perception drive changes in production.
During the teaching intervention, instruction and practice was only given for perception of lexical stress, not for production. If changes were made, it would be direct evidence for the SLM and support previous research that affirm that gains in perception lead to production gains (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Flege, 2003). Testing production differences is an important follow-up study to further explore the perception-production interface.

5.5 Conclusion

The present chapter conceptually replicated Kim (2019) and provided empirical evidence for the positive impact of a teaching intervention in increasing sensitivity to lexical stress. The participants completed four tasks: a perception task, a production task, a week-long teaching intervention, and then repeated the perception task. The results showed a change in response accuracy and sensitivity via d’ between the pre- and post-perception tasks, which provide evidence for the possibility of improving lexical stress perception with a teaching intervention. This improvement provides evidence to support the notion that heritage language classrooms would benefit from instruction to help phonology development. Improvement was not found to be dependent on stress pattern or language, despite only providing instruction in Spanish. For production, duration was the primary cue that children used to denote lexical stress, but this was more evident in paroxytones than in oxytones. As for F0 and intensity, F0 was not used in either language or stress type and intensity was found to
be secondary cue for paroxytones in English and oxytones in Spanish. To compare the perception and production results, age and proficiency were found to be predictors of sensitivity to stress, but not for production of stress. The next chapter will provide a general discussion of all three chapters, along with a summary of the limitations of the current dissertation and future directions for research.
Chapter 6: Discussion

The present work investigated heritage language development with a focus on phonology. This study analyzed the perception and production of lexical stress in Spanish and English in children ages 6-11. The following is a discussion and conclusion to the three experiments that make up this dissertation. I will discuss the findings of the experiments and their relevance with regard to previous research, followed by a brief discussion of the limitations of the dissertation and directions for future research.

6.1 Summary of main findings

Chapter 3 painted a detailed picture of the production of stress of minimal word pairs across two different tasks and two different languages. The first task, a delayed-repetition task, was completed in both English and Spanish. The results revealed that, in both languages, duration is the primary cue that both children and adults use to indicate the tonic syllable of a word. As secondary cues, intensity and F0 seem to have language- and stress-specific differences in the two groups. In English, children use intensity to indicate word stress in paroxytone words but not oxytone words. In Spanish, children, but not adults, use intensity as a secondary cue to indicate V1-V2 differences. With regard to F0, adults, but not children, use F0 to show V1-V2 differences in oxytone words in English and paroxytone words in Spanish. To summarize, whereas children and adults both use duration as the primary means to
producing stress, the children have developed language-specific strategies that differ from the adults with regard to secondary cues. In the context of this study, there were no major changes in production strategies based on age.

Chapter 4 included an AX discrimination task with English- and Spanish-like pseudowords. Pseudowords were chosen instead of real words in order to avoid problems with frequency of lexical tokens. Lexical items that are both highly frequent in both the first person singular present tense and third person singular past tense forms of the verb are not extensive. To disentangle effects of frequency and focus solely on the identification of stress placement in a word, pseudowords that were English-like and Spanish-like were used. The perception task revealed no effect of language. In other words, the participants were able to perceive stress contrasts in English as well as they did in Spanish. The accuracy results also showed that adults have higher accuracy rates than children are on the AX discrimination task. The d’ sensitivity index reflected this difference as well. Adults are much more sensitive to minimal stress pairs than children are. Importantly, age and proficiency were shown to be predictors of sensitivity to stress. It is relevant to note that despite the age and proficiency ranges of the participants in the study, the d’ scores of the children were categorically lower that those of the adults. In other words, the bilingual children studied here (up to 11 years old) were not as sensitive to lexical stress as the adults.

Chapter 5 examined the relationship between perception and production. The purpose of the comparison was to see if there was a discrepancy between the two
modalities in terms of abilities in relation to the adult bilinguals. The results of the production and perception tasks show that the child bilinguals perform more adult-like in their production than in their perception of lexical stress. The perception task showed that the adults were more sensitive than the children to lexical stress, but the children performed above chance in terms of accuracy. In both modalities, there are differences in sensitivity and use of cues when compared to an adult group. In comparison with Chapter 3, this chapter used a smaller subset of data, which resulted in slightly different results in terms of production. In general, they show relatively similar findings with duration being the main cue and intensity as a context-specific secondary true. The third experiment also explored the impact of a teaching intervention on sensitivity to lexical stress. After completing the two tasks, the participants completed a week-long instructional intervention and then repeated the perception task. The results revealed that participants’ sensitivity to lexical stress increased after a week of instruction. Thus, this chapter provided evidence to support the importance of specialized instruction aimed at aiding the development of phonology in the heritage language classroom.

6.2 Relationship between production and perception

In the age range that was tested in the experiment, the children are performing more adult-like in the use of the three cues available to them for production of stress. The results of the production experiments aligned with previous work where production of
L2 stress was related to language experience (proficiency) and age of onset of learning (Au, Oh, Knightly, Jun, & Romo, 2008; Trofimovich & Baker, 2006). As compared to previous research with monolingual children, Hochberg (1988) found that there was some variability in production with children below the age of 5. Here, minimal differences were found between adult and child production of lexical stress.

On the other hand, the children are not as sensitive to lexical stress as adults are. In regard to the perception data, proficiency and age were found to be predictors of sensitivity to lexical stress. This result aligns with previous research as well that found that higher proficiency and language dominance modulate speech perception (Alvord, 2010b, 2010a; Dupoux, Peperkamp, & Sebastián-Gallés, 2010; Guion, Harada, & Clark, 2004; Henriksen, 2012; Kim, 2015, 2019; Rao & Ronquest, 2015). To compare predictors of perception and production, age and proficiency were found to be predictors of sensitivity to stress, but not for production of stress.

### 6.3 A change in comparison groups

A principle aim of this dissertation was to track the course of development of the participants. Previous research on heritage speakers typically includes a comparison group of monolinguals or second language learners. By not including either group, there was conscious effort to move the conversation in heritage language research to one that focuses on a more apt comparison between two groups of bilinguals with similar linguistic experiences. Indeed, this work did include an adult comparison group,
but in contrast with previous research that has been done with heritage speakers, the comparison speakers were adult bilinguals that were members of the same community as the children. The inclusion of this comparison group as opposed to a monolingual or L2 speaker one makes for a more adequate comparison because the input that these speakers receive come from this group of adult speakers. The assumption could be made that the child bilinguals in this study will have similar speech patterns to the adults in the study when they grow up. Ultimately, having this type of comparison instead of one that compares heritage speakers to monolinguals or late L2 learners—one that centers around deficit—is beneficial to understanding heritage language acquisition. A call for a change in comparison groups from native speakers to adult sequential bilinguals is gaining traction in second language acquisition research (see Sakai (2018)).

6.4 Heritage language phonology and teaching

Another important contribution that this dissertation makes to the field of heritage language research concerns education. Traditionally, there is a belief that because of their early exposure there is minimal or no perceptual or production differences for heritage speakers as compared to monolinguals (Boomershine, 2013; Kim, 2015; Ronquest, 2013; Ronquest & Rao, 2018; Willis, 2005). The results of this dissertation support the importance of including phonology instruction in the classroom. Especially, when lexical stress perception poses difficulties for heritage speakers and L2
learners (Beaudrie, 2007, 2012, 2017; Carreira, 2002; Rao & Kuder, 2016). Looking just at the comparison between the adults bilinguals and children, there is a clear discrepancy in that children are much less sensitive to lexical stress than adults are. Furthermore, low proficiency and dominance in English have been shown to negatively impact perception and production in the heritage language (Alvord, 2010b, 2010a; Henriksen, 2012; Rao & Ronquest, 2015). All together, it is obvious that formal instruction in the heritage language is beneficial for speech perception. The current project only explores the impact of instruction with a focus on form via games. It would be prudent to see how other types of instruction impact speech perception. Future studies could compare instruction via games with different types of instruction used in the world language classroom (ex. task-based, inquiry-based, total physical response, focus on meaning, etc).

6.5 Limitations of the study

There are several limitations to the current study. The first limitation deals with the number of participants. Although there were enough participants to reach power for the teaching portion of the study, across all three experiments, the inclusion of a larger participant pool with different groups of heritage learners would have made for interesting comparison. Comparing heritage bilinguals who are currently attending a school where they receive instruction in the heritage language to ones that do not would allow for a more in-depth analysis on the effects of instruction on sensitivity
to stress because factors such as input and use could be added to the analyses.

The pre-/post-test comparison of these two groups could potentially reveal differences in sensitivity just from the general instruction in the heritage language that one group receives that the other does not. This comparison would be especially interesting due to the fact that proficiency was found to be a predictor of sensitivity to stress. The inclusion of a group of heritage speakers who are not attending a dual immersion school, could potentially result in the inclusion of lower proficiency heritage speakers. Having this information in conjunction with the post-test data from both groups would give more strength to the conclusions of the teaching intervention data as it would show how effective the teaching intervention is on a wider range of proficiency and on participants with different amounts of input and use of the heritage language.

The second limitation is concerned with online data collection. Due to the COVID-19 pandemic, some of the data collection procedures were moved online via Zoom. The transfer of experimental procedures to an online platform resulted in a few issues. First, internet connectivity issues were persistent and often times data was lost because the perception task was hosted online and without a stable internet connection. Second, the quality of some of the audio files was suboptimal. Some participants were able to complete the experiment in person and their speech was recorded using professional equipment. Unfortunately, the need to switch to online data collection during the pandemic resulted in some speech samples being recorded via Zoom or mobile phones, which resulted in audio quality that, in some cases, was difficult to
analyze via Praat. Lastly, instruction via Zoom was less than ideal and may have been a factor regarding the effectiveness of the perception games. In-person instruction with small groups of participants would have assured that I was able to assess the full effects of instructional game play in comparison with individualized online instruction.

A third limitation concerns the use of the language assessment as a proxy for proficiency. Although the school district from which I was collecting data used the Test of Phonological Awareness (TOPA-2+) for their own purposes to assess students and tailor instruction, to my knowledge it has not been used previously in heritage language or L2 research with children. For this reason the external validity of the measure is uncertain. Future investigations should explore the use of the TOPA-2+ in novel linguistic research. Regardless of the external validity, the fact remains that TOPA-2+ scores did show strong correlations with sensitivity to lexical stress.

6.6 Directions for future research

Based on the findings of the current study, several suggestions can be made for future research. First, for direct evidence to support the SLM’s claims that perception drives production, it would be prudent to replicate the current study and add a production task to the post-teaching intervention tasks. The pedagogical intervention of instructional games in the third experiment in this dissertation focused solely on improving perception skills. The results of the third experiment showed that after the hour of
instruction there was an improvement in perception skills, but production skills were not measured after the teaching intervention. Unfortunately, the dissertation did not explore this relationship further by including a post-teaching intervention production task to see if perception gains led to production gains. This would have been direct support to the notion that gains in perception lead to gains in production (Akahane-Yamada, Tohkura, Bradlow, & Pisoni, 1996; Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Flege, 2003; Flege, Bohn, & Jang, 1997). In a follow-up study, a second delayed-repetition task would be prudent to add for a better analysis of the relationship between perception and production.

A second direction of research that should be considered for a follow-up study is the addition of a delayed post-test. Even though the results showed perception gains, the post-test was completed shortly after a week-long instruction. In a future study, we could give a delayed post-test to the students a month after instruction. This delayed post-test would give a better sense of how well the participants retained the gains in perception.

6.7 Conclusion

This dissertation began with the goal of studying the development of heritage language phonology. The cross-sectional data suggest that proficiency is an important predictor to mediating sensitivity to perceiving stress contrasts, but not producing them. Importantly, pedagogical interventions in the language classroom that are
aimed at improving perception skills are important to increasing sensitivity to stress contrasts. Overall, this dissertation adds valuable information to heritage language and speech perception research. Together, the three experiments in this dissertation contribute to our understanding of speech learning models in that they add information about the acquisition of perception of suprasegmentals. It is an important step in the right direction to further explore in the field of heritage language phonology, and for tracking language development in heritage bilingual children in general. Specifically, it aides in shifting the conversation around heritage speakers away from the focus and comparison between monolinguals and second language learners, to a comparison with adult bilinguals or other heritage language learners.
Appendix A: Bilingual Language Profile

1. What is your full name?

2. How old are you?

3. Current grade?

4. Where is your family from?

5. At what age did you start learning the following languages?
   - English
   - Spanish

6. At what age did you start to feel comfortable using the following languages?
   - English
   - Spanish

7. In an average week, what percentage of the time do you use the following languages with friends?
   - English
   - Spanish

8. In an average week, what percentage of the time do you use the following languages with family?
   - English
   - Spanish

9. In an average week, what percentage of the time do you use the following languages at school?
English

Spanish

10. How well do you speak English?
   1 - 2 - 3 - 4 - 5 - 6

11. How well do you speak Spanish?
   1 - 2 - 3 - 4 - 5 - 6

12. How well do you understand English?
   1 - 2 - 3 - 4 - 5 - 6

13. How well do you understand Spanish?
   1 - 2 - 3 - 4 - 5 - 6

14. How well do you read English?
   1 - 2 - 3 - 4 - 5 - 6

15. How well do you read Spanish?
   1 - 2 - 3 - 4 - 5 - 6

16. How well do you write English?
   1 - 2 - 3 - 4 - 5 - 6

17. How well do you write Spanish?
   1 - 2 - 3 - 4 - 5 - 6

18. I identify with an English-speaking culture.
   1 - 2 - 3 - 4 - 5 - 6

19. I identify with a Spanish-speaking culture
Appendix B: Test of Phonological Awareness (TOPA-2+)

Examples of the final sound identification tasks:

1. Say rose without the /s/.
2. Say train without the /n/.
3. Say group without the /p/.
4. Say seat without the /t/.
5. Say bake without the /k/.
6. Say inch without the /ch/.
7. Tell me the last sound in the word “pick.”
8. Tell me the last sound in the word “ran.”
9. Tell me the last sound in the word “fill.”
10. Tell me the last sound in the word “bug.”
11. Tell me the last sound in the word “same.”
12. Tell me the last sound in the word “tooth.”
Appendix C: Delayed-Repetition Task

Below is the list of the words used in the delayed-repetition task for English and Spanish.

**English:**

1. She won the *contest* yesterday.
2. I *contest* the results.
3. There is a *present* here.
4. I have to *present* today.
5. We buy *produce* at the store.
6. We *produce* documents.
7. I *address* the crowd.
8. My *address* is a secret.
9. We have a *project* today.
10. The words *project* on tv.
11. The *content* of the paper.
12. We are *content* today.
13. There is a *conflict* in class.
14. Our schedules *conflict* in school.
15. We *object* to the rule.
16. The *object* on the table.
Spanish:

1. Yo *canto* contigo. (Eng. *I sing with you*)
2. Ayer *cantó* el niño. (Eng. *Yesterday the boy sang*)
3. Hoy *hablo* con mi amiga. (Eng. *Today I talk with my friend*)
4. Mi mamá *habló* contigo. (Eng. *My mom talks to you*)
5. En clase *grito* la respuesta. (Eng. *In class I shout the answer*)
6. La maestra *gritó* ayer. (Eng. *La teacher shouted yesterday*)
7. Yo *dibujo* en casa. (Eng. *I draw at home*)
8. Ella *dibujó* por la mañana. (Eng. *She drew in the morning*)
9. Yo *camino* con el perro. (Eng. *I walk with the dog*)
10. El chico *caminó* anoche. (Eng. *The boy walked last night*)
11. Hoy *trabajo* contigo. (Eng. *Today I work with you*)
13. Yo *escucho* la canción. (Eng. *I listen to the song*)
14. El niño *escuchó* la canción. (Eng. *The boy listened to the song*)
15. Yo *pinto* una pintura. (Eng. *I paint the picture*)
16. Ella *pintó* en clase. (Eng. *She painted in class*)
Appendix D: Elicited Production Task

This appendix includes the picture prompt stimuli used in the elicited production task.
Figure D.1: The images used to represent target words in the elicited production task.

(a) Caminar
   to walk

(b) Cantar
   to sing

(c) Dibujar
   to draw

(d) Trabajar
   to work

(e) Escuchar
   to listen

(f) Hablar
   to talk

(g) Gritar
   to shout

(h) Pintar
   to paint
Appendix E: AX discrimination Task

Below is a list of the English-like and Spanish-like pseudowords used in the AX discrimination task. Participants heard all stimuli with ultimate and penultimate stress.

**English:**

1. precrack
2. sestrow
3. strorpez
4. formand
5. fonnain
6. ronvoon
7. torvoot
8. bendict
9. brendict
10. bontoon

**Spanish:**

1. restin
2. desbo
3. resi
4. reptin
5. testus
6. bonte
7. sesfon
8. repa
9. rastis
10. salpen
Appendix F: Posterior predictive checks for statistical models fit to production data in Chapter 3

This appendix includes posterior predictive checks for the 5 models fit to the production data presented in Chapter 3. In each case the outcome variable was a z-score (standard score) for the production metrics (duration, intensity, pitch). The models plotted in Figure F.1 were the following:

1. b-prod-mod-00: z-score as a function of lexical stress in English and Spanish (children).
2. b-prod-mod-00-adults: z-score as a function of lexical stress in English and Spanish (adults).
3. b-prod-mod-01: z-score as a function of lexical stress and age for English and Spanish.
4. b-prod-mod-02: z-score as a function of task type.
5. b-prod-mod-03: z-score as a function of lexical stress, metric and language.

For detailed information regarding the nature of the generative model specified in each case see section 3.4.
Figure F.1: Posterior predictive checks for generative models fit to production data.
Appendix G: Posterior predictive checks for statistical models fit to perception data in Chapter 4

This appendix includes posterior predictive checks for the 4 models fit to the perception data presented in Chapter 4. In each case the outcome variable was a z-score (standard score) for correct response or d’ score. The models plotted in Figure G.1 were the following:

1. b-perc-mod-00: z-score of correct responses as a function of language (children).
2. b-perc-mod-00-adults: z-score of correct responses as a function of language (adults).
3. b-perc-mod-01: z-score of sensitivity as a function of language.
4. b-perc-mod-03: z-score of sensitivity as a function of proficiency, age, and language.

For detailed information regarding the nature of the generative model specified in each case see section 4.3.
Figure G.1: Posterior predictive checks for generative models fit to perception data.
Appendix H: Posterior predictive checks for statistical models fit to perception data in Chapter 5

This appendix includes posterior predictive checks for the 3 models fit to the perception data presented in Chapter 5. In each case the outcome variable was a z-score (standard score) for correct responses and sensitivity to stress. The models plotted in Figure H.1 were the following:

1. b-perc-time-mod-00: z-score of correct responses as a function of time in English and Spanish.
2. b-perc-time-mod-01: z-score of sensitivity as a function of time in English and Spanish (adults).
3. b-perc-time-mod-02: z-score of change in sensitivity as a function of language, age, and proficiency.

For detailed information regarding the nature of the generative model specified in each case see section 5.3.
Figure H.1: Posterior predictive checks for generative models fit to perception data in pre- and post-tests.
Appendix I: Power Analysis for AX discrimination - pedagogical intervention

In this appendix I describe the power analysis used to plan the pre/post AX discrimination experiment. As a starting point, I used the mean and standard deviation d’ values of the heritage speaker group performance in the discrimination task presented in Kim (2019) (Mean: 2.41, SD: 0.73) Following McAndrews (2019), which explores numerous studies including pedagogical interventions, I anticipated a small to moderate effect of a difference in means of about 0.3 (d’). I set alpha at 0.05 and calculated the number of participants necessary to obtain 0.8 power. The following code uses this information to run the power analysis.

```r
# Settings
power <- 0.8 # standard
time1_mean <- 2.41 # from Kim (2019)
time2_mean <- 2.71 # ES of 0.3
delta <- time2_mean - time1_mean
est_sd <- 0.73 # from Kim (2019)
alpha <- 0.05 # standard false positive rate of 5%
```

The `power.t.test` function takes the above information as arguments and runs the test.
Assuming the pedagogical intervention works and produces a mean difference in sensitivity scores (d’) of 0.3, the power analysis suggests the design has an 80% chance of finding the effect with 38 participants.
REFERENCES


Boomershine, A. (2013). The perception of English vowels by monolingual, bilingual, and heritage speakers of Spanish and English. In S. E. B. Howe Chad & M. L. Quesada (Eds.), *Selected proceedings of the 15th hispanic linguistics symposium* (pp. 103–118).


Face, Timothy L. (2006). Cognitive factors in the perception of Spanish stress
http://doi.org/10.1515/LING.2006.040


Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds...
"L" and "R". *Neuropsychologia*, 9, 317–323.


SC.


Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identify-


Kuhl, P. K., & Miller, J. D. (1978). Speech perception by the chinchilla: Iden-


of training and a study abroad program for L1 English late learners of Spanish.


111–114.


Stan Development Team. (2018). Stan modeling language users guide and ref-


Simon Fraser University, Burnaby, British Columbia, Canada.


