(DIS)CONNECTING PERCEPTION AND PRODUCTION: TRAINING ADULT NATIVE SPEAKERS OF SPANISH ON THE ENGLISH /i/-/ɪ/ DISTINCTION

A Dissertation
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Linguistics

By

Mari Sakai, M.A.T.

Washington, D.C.
October 13, 2016
This dissertation features three experiments that investigated how perception and production are connected in the acquisition of second language (L2) phones by comparing the effectiveness of two modality-specific trainings and their respective potential for cross-modality gains. Participants were native speakers of Spanish with advanced English proficiency, and the targets were the English vowels /i/ and /ɪ/.

In Experiment 1, participants ($n=15$) received perception-only training; they heard auditory exemplars of the target phonemes but never produced the sounds. In Experiment 2, two variations of a production-only training were compared that either allowed or denied access to the auditory feedback loop. A first group ($n=14$) underwent training using a computer program that provided real-time visual representations of spoken vowels. They never heard any other-produced auditory tokens of the target sounds, although they could hear the sound of their own voices. A second group ($n=15$) underwent the same training, but wore noise-cancelling headphones and listened to white noise. This ensured that they never heard other- or self-generated tokens of the target phonemes, for the first time in the literature truly isolating production from all auditory influence. All participants in both experiments completed a battery of pre- and post-tests in perception and production, and they were also compared against a control group ($n=15$) and two baselines: a group of native speakers of English ($n=20$), and a bilingual group ($n=16$) who was deemed to have acquired /i/ and /ɪ/. In Experiment 3, the two
baselines were directly compared in order to test their efficacy as benchmarks for phonetic training experiments.

Results revealed that: (1) perception-only training led to large gains in perception and no sizeable improvements in production; (2) production-only training led to variable results for production, and medium-sized improvements in perception; (3) access to the auditory feedback loop provided a benefit to production; (4) access to or denial of the auditory feedback loop did not affect cross-modal learning in perception; and (5) bilinguals are a fitting, and for many purposes likely sufficient, comparison baseline group in L2 speech training experiments.

The dissertation contributes novel theoretical, methodological, and educational insights to the L2 speech training literature.
ACKNOWLEDGMENTS

The past five years, being a Ph.D. student at Georgetown, have been some of the best of my life. I have learned and grown in so many ways, and I have many people to thank. First and foremost, it has been a pleasure and an honor to work with my committee, Lourdes Ortega, Tracey Derwing, and Jennifer Nycz. Each of them has guided me along my doctoral path in various ways. Lourdes is not only brilliant, but also kind. I have learned from her expertise and the expansive breadth of her knowledge. She is so generous with her time, and she has helped me to see my own research with new eyes. She is truly the best mentor and advisor any Ph.D. student could wish for. It was because of Lourdes’ suggestion to attend a pronunciation conference that I met Tracey Derwing. In my very first conversation with Tracey at PSLLT in Ames, Iowa, she made an off-hand suggestion that would propel my dissertation idea into an important study. Our second meeting at PSLLT the next year ended in an hour-long conversation that covered academic and personal topics, and it was in this chat that Tracey agreed to join my committee. I feel so fortunate to have had her as a mentor these past few years. Like Lourdes, she effortlessly blends expertise in academia with a kind-hearted personality. My third committee member, Jen Nycz, is an example of an incredibly intelligent phonetician, who has already achieved so much in her career. I was her teaching assistant for a semester and was continuously in awe of the depth of her knowledge. These three women are my role models, and I will forever appreciate the foundation they have built for me.

I also need to thank my family and friends. My parents and sisters are the most important people in my life. They are intertwined in my identity, and they are who I am. My mom and dad have made my success possible, as they have supported and pushed me at every step of my life
and education. My mom has always been my example of an intelligent, compassionate, and hard working person. Because she taught at my elementary school, school has always felt like home. My father has given me an example of a strong, diligent, and hard working person, and he has passed on those expectations to me. There are no excuses; you just do it. At my moment of maximum stress in the dissertation process, they flew to D.C. to rescue me just as I was breaking down from starting a new job, moving apartments, and analyzing data all at the same time. They literally gave me time and rejuvenated me emotionally to be able to keep going.

Finally, I need to show my greatest appreciation for my friends who have supported me through this process. John Chi, Lauren Park, Karen Feagin, Marta Baffy, Julie Lake, Shaenelle Wilson, Claudine Wilson, Jenifer Jesse, Tonya Fong-Yee, Qiana Lee, and Miche Nguyen…they literally fed me and took care of me during the hardest dissertation writing times. Thank you to my “older siblings” at Georgetown: Luke Amoroso, Sarah Young, Phillip Hamrick, and Yoonsang Song. Thank you to the one and only member of my Applied cohort, Joshua Kraut, who tinted my time at Georgetown with warmth and laughter. A huge hug to the Georgetown Phonology Contingent, Charles Nagle and Germán Zárate-Sánchez. I couldn’t be happier that we reached ultimate attainment in Denmark, and I am so excited to continue our professional journeys together. To my Georgetown friends who I haven’t mentioned yet, Mackenzie Price, Young A Son, and Heidi Getz, and everyone who I’ve mentioned above: You have made these five years incredible. I am so lucky to have made such amazing friends.

I never expected that coming to Georgetown would have given me all of this. These five years of my life have been incredible, and I owe so much of it to you all. Thank you for supporting me and helping to make this journey so meaningful.
# TABLE OF CONTENTS

Chapter 1: Introduction ........................................................................................................... 1

Chapter 2: Understanding Speech Perception and Production through the Lens of Second Language Acquisition .................................................................................................................. 5
  2.1. What theory predicts ........................................................................................................... 5
    2.1.1. The Speech Learning Model .......................................................................................... 5
  2.2. Behavioral evidence ........................................................................................................... 7
    2.2.1. Correlation studies ....................................................................................................... 7
    2.2.2. Perception training ...................................................................................................... 14
    2.2.3. Production training ...................................................................................................... 19
  2.3. Second language vowel development of native speakers of Spanish learning English ..... 25

Chapter 3: Interdisciplinary Contributions to the Perception-Production Conversation ........... 28
  3.1. Deaf and hard-of-hearing populations .............................................................................. 28
  3.2. Neurocognition ................................................................................................................ 32

Chapter 4: Research Questions ................................................................................................ 37

Chapter 5: Method .................................................................................................................... 42
  5.1. Participants ....................................................................................................................... 42
  5.2. Materials .......................................................................................................................... 47
    5.2.1. Target vowels and auditory stimuli ............................................................................. 47
    5.2.2. Pre- and post-tests ..................................................................................................... 50
    5.2.3. Native speaker identifications of productions ............................................................. 56
    5.2.4. Other data collection instruments .............................................................................. 57
  5.3. Coding and analysis of test performances ....................................................................... 58
5.4. Experiment 1 .................................................................................................................. 61
  5.4.1. Perception-only training ......................................................................................... 62
  5.4.2. Procedure .............................................................................................................. 63
5.5. Experiment 2 .................................................................................................................. 64
  5.5.1. Production-only training task ................................................................................. 65
  5.5.2. Procedure .............................................................................................................. 71
5.6. Experiment 3 .................................................................................................................. 72
  5.6.1. Procedures ............................................................................................................ 72
  5.6.2. Coding and analysis ............................................................................................. 72

Chapter 6: Results ................................................................................................................. 74
6.1. Experiment 1 .................................................................................................................. 74
  6.1.1. Research Question 1: Does perception-only training of two second language vowels improve perception of the same sounds? ................................................................. 74
  6.1.2. Research Question 2: Does perception-only training of two second language vowels improve the production of the same sounds? ........................................................................... 84
6.2. Experiment 2 .................................................................................................................. 94
  6.2.1. Research Question 1: Does production-only training of two second language vowels improve production of the same sounds? ........................................................................... 95
  6.2.2. Research Question 2: Does production-only training improve the perception of two target second language vowels? .................................................................................. 109
  6.2.3. Research Question 3: Does access to the auditory feedback loop (for the Production-Only group) versus no access (for the Production-Only, No Sound group) influence the
development of second language speech perception and production after production-only training? ........................................................................................................................................ 117

6.3. Experiment 3 ........................................................................................................................................................................... 121

6.3.1. Research Question 1: How comparable and useful are bilingual and native speaker baseline groups? ........................................................................................................................................ 121

Chapter 7: Discussion and Conclusion ........................................................................................................................................ 135

7.1. Summary of the main results .................................................................................................................................................. 135

7.2. Speech training leads to same-modality gains .......................................................................................................................... 140

7.3. Speech training can support cross-modal benefits ................................................................................................................. 145

7.4. Understanding production to perception learning .................................................................................................................. 149

7.5. Do the findings support the Speech Learning Model? ........................................................................................................... 151

7.6. A look at advanced learners who are motivated to overcome pronunciation difficulties. 153

7.7. What baselines do we want for speech training studies? ......................................................................................................... 161

7.8. Methodological lessons learned .............................................................................................................................................. 165

7.9. Study limitations ....................................................................................................................................................................... 169

7.10. Pedagogical implications ......................................................................................................................................................... 173

Appendices .................................................................................................................................................................................... 175

Appendix A. Recruitment flyer .................................................................................................................................................... 175

Appendix B. Stimuli words and images ........................................................................................................................................ 176

Appendix C. Illustrations used for a production task .................................................................................................................. 180

Appendix D. Embedded words in a production task .................................................................................................................. 181

Appendix E. Passages used in a production task ......................................................................................................................... 182

Appendix F. Instructions for the production training program .................................................................................................. 184
Appendix G. Articulatory information used in the production training ........................................ 186

Appendix H. Compiled scores for all groups ........................................................................ 187

References .................................................................................................................................. 188
LIST OF FIGURES

Figure 1. Within-group effect size calculations by manner of articulation ........................................ 18

Figure 2. Grey boxes represent all of the synthesized vowel tokens that were created along spectral and duration continuums ................................................................. 49

Figure 3. Blue boxes represent the synthesized tokens that NS of English identified as /i/ sounds, and pink boxes represent the tokens identified as /ɪ/ sounds ........................................... 59

Figure 4. The Vowel Shapes visualizer ................................................................................................. 66

Figure 5. Representation of the two laptop screens participants saw during the production-only training program ........................................................................................................ 68

Figure 6. Identification pre- and post-test scores for the experimental Perception-Only (Per) and Control (C) groups .................................................................................................... 77

Figure 7. Discrimination task scores across spectral steps for the Perception-Only (Per), Control (C), NS English (E), and Bilingual Acquired (B) groups. Graph (A) depicts pre-test scores, and (B) depicts post-test scores ........................................................ 82

Figure 8. Discrimination task scores across duration steps for the Perception-Only (Per), Control (C), NS English (E), and Bilingual Acquired (B) groups ........................................................ 83

Figure 9. Production pre- and post-test scores for the Perception-Only (Per) and Control (C) groups on all three tasks ........................................................................................................ 87

Figure 10. Z-score plots of the Trained group’s productions on all three tasks at pre- (A) and post-test (B) ..................................................................................................................... 88

Figure 11. Z-score plots of the Control group’s productions from all three tasks at pre-test (A) and post-test (B) ............................................................................................................. 89

Figure 12. Bark normalized mean acoustic measurements for /i/ and /ɪ/ productions on the Elicited Production task for the Perception-Only (Per), Control (C), NS English, and Bilingual Acquired (B) groups ................................................................. 92

Figure 13. Bark normalized mean acoustic measurements for /i/ and /ɪ/ productions on the Passage Reading task ........................................................................................................... 93

Figure 14. Bark normalized mean acoustic measurements for /i/ and /ɪ/ productions on the Picture Description task ........................................................................................................... 93

Figure 15. Production pre- and post-test scores for the Production-Only (Pro), Production-Only, No Sound (No Sound), and Control (C) groups on all three tasks ........................................ 97
Figure 16. Z-score plots of the Production-Only group’s productions on all three tasks at pre-test 
(A) and post-test (B).................................................................................................................. 104

Figure 17. Z-score plots of the Production-Only, No Sound group’s productions on all three tasks 
at pre-test (A) and post-test (B)........................................................................................................ 105

Figure 18. Bark normalized mean /i/ and /ɪ/ productions for the control and two trained groups on 
the Elicited Production task.............................................................................................................. 108

Figure 19. Bark normalized mean /i/ and /ɪ/ productions for the control and two trained groups on 
the Passage Reading task.................................................................................................................. 108

Figure 20. Bark normalized mean /i/ and /ɪ/ productions for the control and two trained groups on 
the Picture Description task........................................................................................................... 109

Figure 21. Identification task pre- and post-test scores for the two experimental one control 
group on each of the three sets of stimuli........................................................................................ 111

Figure 22. Discrimination task scores across spectral steps for the Production-Only (Pro), 
Production-Only, No Sound (Pro, NS), Control (C), NS English (E), and Bilingual Acquired 
(B) groups at pre-test (Graph A) and post-test (Graph B)................................................................ 114

Figure 23. Discrimination task scores across duration steps for the Production-Only (Pro), 
Production-Only, No Sound (Pro, NS), Control (C), NS English (E), and Bilingual Acquired 
(B) groups at pre-test (Graph A) and post-test (Graph B).................................................................. 115

Figure 24. NS English and Bilingual Acquired scores on the Identification task with three types 
of stimuli: (A) shVp, (B) fVt, and (C) naturally recorded tokens ......................................................... 123

Figure 25. Bilingual Acquired and NS English group performance on the Discrimination task 125

Figure 26. NS English and Bilingual Acquired groups’ scores on three production tasks: (A) 
Elicited Production, (B) Passage Reading, and (C) Picture Description........................................ 127

Figure 27. Graph A depicts the Bilingual Acquired group’s /i/ and /ɪ/ vowels and B presents the 
vowels for the NS English group ...................................................................................................... 130

Figure 28. Average /i/ and /ɪ/ vowel productions by NS English and Bilingual Acquired groups 
on the Elicited Production task........................................................................................................ 131

Figure 29. Average /i/ and /ɪ/ vowel productions by NS English and Bilingual Acquired groups 
on the Passage Reading task............................................................................................................ 132

Figure 30. Average /i/ and /ɪ/ vowel productions by NS English and Bilingual Acquired groups 
on the Picture Description task........................................................................................................ 132

Figure 31. Pie charts that represent the percentage of participants who had high scores in each of 
the modalities at the prescreen-pre-test stage.................................................................................. 158
LIST OF TABLES

Table 1. Sample of perception and production correlation studies .......................................................... 9
Table 2. Summary of results from Goto (1971) and Sheldon & Strange (1982) ........................................ 12
Table 3. Within-group Cohen's $d$ effect size scores for eighteen studies, with means (SD) presented in the last row of each column (from Sakai & Moorman, submitted) .............................. 16
Table 4. Regions of the brain associated with perceiving and producing speech (summary of a portion of Table 2 in Price, 2012) .................................................................................. 33
Table 5. Participant demographic and linguistic background information ................................................ 43
Table 6. Description of testing and training days for the Perception-Only experimental group ................ 64
Table 7. Description of trials and blocks for each day of production training ........................................... 71
Table 8. Comparability of features in the perception-only training of Experiment 1 and the production-only training of Experiment 2 .................................................................................. 71
Table 9. Mean percent accuracy (SD) for three types of stimuli in the Identification task for Perception-Only, Control, and baseline groups ........................................................................... 76
Table 10. Mean percent accuracy (SD) for the Perception-Only training group for the two trained vowels on each day of training ..................................................................................... 76
Table 11. Mean (SD) native speaker identifications for three production tasks for the Perception-Only, Control, and baseline groups ................................................................................. 86
Table 12. Mean (SD) of percent overlap of /i/ and /i/ for the Perception-Only, Control, and baseline groups .......................................................................................................................... 90
Table 13. Z-score coordinates ($Z_3-Z_2$, $Z_3-Z_1$) for /i/ (A) and /i/ (B) on the three production tasks at pre- and post-test ........................................................................................................ 92
Table 14. Mean (SD) of native speaker identifications for three production tasks for the two experimental, one control, and two baseline groups ....................................................................... 96
Table 15. Mean percent accuracy score (SD) for two Production-Only training groups of the two vowels on each day of training .......................................................................................... 96
Table 16. Mean (SD) of percent overlap of the /i/ and /i/ categories for all experimental, control, and baseline groups at pre- and post-test ........................................................................ 106
Table 17. Coordinates in $z$-scores ($Z_3-Z_2$, $Z_3-Z_1$) for /i/ (A) and /i/ (B) in all three tasks at pre- and post-test .................................................................................................................... 106
Table 18. Mean (SD) for three sets of stimuli on the Identification task at pre- and post-test... 110

Table 19. A comparison of the Production-Only and Production-Only, No Sound training groups' performance on perception and production tasks............................................................................ 118

Table 20. Bilingual Acquired participants who scored lower than the NS English range on the perception tasks or production tasks ............................................................................................................. 128

Table 21. Within group (pre-to-post) Cohen's d effect sizes and p-values from paired samples t-tests for the three trained groups on the perception and production tasks ......................................... 136

Table 22. Mean scores on the Identification task with three types of stimuli for the participants in the Production-High group.............................................................................................................. 154
Chapter 1: Introduction

One of the challenges that second language learners face is the acquisition of new sounds, both the creation of accurate mental representations of phonemes and the ability to articulate them. Learning a nonnative sound system is known to pose great challenges, and attaining levels of pronunciation in the second language (L2) that are native-like is considered by many to be practically impossible (e.g., Hyltenstam & Abrahamsson, 2000; Long, 2013), although some dissenting voices remain (e.g., Ioup, Boustagui, Tigi, & Moselle, 1994; Moyer, 2014). Many factors have been shown to contribute to the degree of ultimate success of second language pronunciation, including age of L2 onset (e.g., Long, 1990), age of arrival (e.g., Flege, Munro, & MacKay, 1995), first language (e.g., Derwing & Munro, 2013), continued L1 use (e.g., Piske, MacKay, & Flege, 2001), and motivation (e.g., Moyer, 1999). Nevertheless, Flege's (1995, 2003) Speech Learning Model (SLM) states that adults do retain the ability to create new phonetic categories for second language sounds. A second postulate of the SLM states that a learner must first achieve target-like perception in order to produce the phonemes accurately. This theory of language learning, therefore, places a large emphasis on the interconnectedness of the perception and production modalities and makes perception a developmental prerequisite of production.

In this dissertation, I attempted to separate the processes that underlie accurate perception and production acquisition of two English phonemes, /i/-/ɪ/, in order to better understand how the two modalities are connected. Drawing from methodology in phonetic training (i.e., Kondaurova & Francis, 2010) and speech visualization (i.e., Ryan, Madsen, Buckley, Alex, & Bronstein, 2014), I report on two interrelated experiments that trained native speakers of Spanish in perceiving (Experiment 1) or producing (Experiment 2) this phonemic contrast, which is a
known difficult contrast for this population of learners. I also report on the comparison of the perception and production abilities of two baseline groups (Experiment 3), one representing the traditional native-speaking benchmark used in most L2 studies to date, and another representing a bilingual baseline of L1 Spanish–L2 English users who were deemed to have acquired the phonemic /i/-/ɪ/ distinction.

In the first two experiments, I used second language phonetic acquisition as a means to explore the overlap of the perception and production modes. Experiment 1 is a perception training program in the absence of production practice. This type of training has an established tradition in the L2 phonology literature, and I inspected evidence that perception training aids perception but also whether it may aid production. Experiment 2 is a production training program that offered participants no aural input. The second experiment included a training group that wore noise-cancelling headphones and listened to pink or brown noise in order to eliminate self-produced aural feedback. To the best of my knowledge, this is the first study that truly isolated perceptual influence in a production training experiment. The goal was to ascertain whether production training aids production and, in a mirror image of Experiment 1, whether it may aid perception as well. In both experiments, participants completed perception and production pre- and post-tests and were compared to three groups: (1) a control group, who completed the pre- and post-tests without any training; (2) a functionally monolingual English native-speaking group; and (3) a Spanish-English bilingual group, who showed evidence of having acquired the two vowel categories. The native speakers of English and the bilingual group were baselines and only completed the pre-test battery. Finally, the direct comparison of the two baselines in Experiment 3 sought to determine the usefulness of establishing
performance yardsticks in L2 phonetic development by resorting to a native speaker or a bilingual target in training studies.

The dissertation is organized as follows. In Chapter 2, I first outline the contributions second language research has added to the debate of the connected nature of speech perception and production. After presenting an overview of Flege’s (1995, 2003) Speech Learning Model, I examine behavioral evidence from three bodies of literature within second language acquisition (SLA): (1) studies that have run correlational analyses between scores on perception and production tasks, (2) studies featuring perception training that have tested for gains in production, and (3) studies featuring production training that have tested for gains in perception. I also briefly discuss what is known about the acquisition of L2 English vowels by L1 Spanish speakers. Next, in Chapter 3, I explore contributions from two additional fields, the deaf and hard-of-hearing population and neurocognitive sciences, which offer an interdisciplinary perspective. Studies of people with profound hearing loss can reveal associations between perception and production, as this population acquires language in the absence of complete auditory information. Neurocognitive evidence addresses areas of the brain that are activated for both perception and production tasks in optimal and compromised environments. In the subsequent two chapters, I present my research questions (Chapter 4) and describe the three experiments I conducted for the dissertation (Chapter 5). The results for each of the three experiments are presented in Chapter 6. The dissertation ends with an extended discussion of the findings, and what they mean in terms of theory and research methodology. I close the chapter with an acknowledgement of the limitations of the study and a reflection of the pedagogical implications.
The dissertation contributes to the body of literature that attempts to understand the complex association of speech perception and production processing, and its influence on second language phonetic acquisition.
Chapter 2: Understanding Speech Perception and Production through the Lens of Second Language Acquisition

In this chapter I first present an overview of Flege’s (1995, 2003) Speech Learning Model. I then examine behavioral evidence from three bodies of literature within second language acquisition (SLA): (1) studies that have run correlation analyses between scores on perception and production tasks, (2) studies featuring perception training that have tested for gains in production, and (3) studies featuring production training that have tested for gains in perception. I close the chapter with a discussion of what is known about the acquisition of L2 English vowels by L1 Spanish speakers.

2.1. What theory predicts

2.1.1. The Speech Learning Model

In a series of publications spanning almost two decades, Flege outlined a theory of second language speech acquisition called the Speech Learning Model (Flege, 1988, 1992, 1995, 1999, 2002, 2003). He explains that the model begins with two broad assumptions: (1) the bilingual’s L1 and L2 phonetic subsystems are not separable, as they both exist in the same long-term memory space, and (2) the capacity to create new phonetic categories remains intact throughout the lifetime (Flege, 2003). To expand on the first assumption, when adults begin learning a second language, they already have an L1 phonemic inventory intact. In the process of acquiring the non-native phonemic system, learners undergo a period where their first language categories mitigate the acquisition of the L2 phonemes. As Cenoz and García Lecumberri (1999) explain, adult learners “rely on phonetic rather than sensory perception so that their perception of L2 sounds is biased by their L1 phonetic system and [learners] tend to perceive L2 sounds in
terms of the categories in the L1” (p. 262). In cases where a phone from the L2 is discernable as a distinct sound from the L1 phonemes, a new mental representation will successfully be added to the native phonemic system. However, in instances when a learner cannot discriminate differential features of an L2 phone, the sound assimilates onto an already existing L1 representation to create a composite category, which carries information of both the L1 and the L2 phones. Flege’s SLM hypothesizes that once learners are able to discern at least one differential feature of the L2 sound, they will be able to create a new L2 category.

Flege’s first assumption, that there is flexibility in expanding and creating new categories, is supported by numerous perception training studies that show learners can improve to native-like levels in the discrimination and identification of L2 contrasts that were initially difficult to perceive (e.g. Kondaurova & Francis, 2010; Iverson & Evans, 2009). The plasticity of phonetic categories found in adult learners substantiates Flege’s second assumption that the ability to create new categories lasts throughout adulthood.

A subsequent third hypothesis of the SLM extends to the production modality and states that non-native speakers’ productions will eventually reflect the mental representations of the L2 phonemes (Flege, 1995, 2003). This hypothesis rests on the assumption, discussed in Flege (1995), that the mental representation consists of phonetic categories that exist in long-term memory and carry both acoustic and articulatory information. Flege (2003) further explains that there is an intricate and detailed auditory-articulatory mapping that connects what language users hear to what they produce. If the mental representation is fine-tuned by accurately perceiving distinctive features of the L2 sound, and production mirrors the mental representation, then production is constrained by perceptual abilities. Flege asserts, “L2 phonetic segments cannot be produced accurately unless they are perceived accurately” (2003, p. 27). This claim implies that
a non-native speaker’s production accuracy will not precede the perceptual accuracy. Furthermore, it is not expected that productive abilities will “lead” or be more native-like than segmental perception (p. 26).

Although the SLM contains many postulates and hypotheses, to summarize, the claims relevant for this dissertation are the following (Flege, 2003):

1. Perception of L2 segments can improve after sufficient exposure.

2. Accurate perception necessarily precedes accurate production, thus, perception and production are correlated, but their connection is unidirectional, from perception to production.

These components of the SLM indicate that there is a strong connection between the perceptive and productive systems in terms of second language phonetic acquisition. Next, I will review SLA literature that support, complicate, and challenge these claims.

2.2. Behavioral evidence

The past forty years of SLA literature shows evidence of a strong interest in the connection between learners’ perception and production abilities. In this time, hundreds of researchers have sought to determine if there is a relationship between these modalities by using many behavioral tasks, including discrimination, identification, acoustic analysis, and training programs. First I will explore studies that have correlated scores on perceptive and productive tasks, then I will turn to training programs in depth.

2.2.1. Correlation studies

One of the simplest analyses to reveal a relationship between perception and production is a correlation. The SLM posits that a learner will produce a second language phone correctly only if he or she has already mastered accurate perception. This predicts a higher production
score will be associated with a higher perception accuracy score, revealing a positive correlation. A sample of SLA literature that has run correlational analyses on perception and production measures reveals disparate results and indicates that discrepancies might be a result of several methodological and substantive factors, including sample size, target segments, first and second language combinations, and tasks. Before I review correlation studies, two complications must be addressed first.

In different publications, Flege describes the expected correlation of perception and production scores in two different ways. First, as was quoted in the previous section, he says, “L2 phonetic segments cannot be produced accurately unless they are perceived accurately” (2003, p. 27). However, in a different publication that reported findings from a study on Italian speakers’ perceptions and productions of English vowels, Flege, MacKay, and Meador (1999) stated that a “hypothesis of the speech learning model, viz., that the accuracy with which L2 vowels are produced is limited by how accurately they are perceived” (p. 2973). The first interpretation of the SLM hypothesis from Flege (2003) suggests that there is some sort of threshold of accurate perception that must be reached before accurate production can be achieved. This depicts an asymptotic curve that can be found by running a regression. On the other hand, the second quote from Flege et al. (1999) implies that there is a continuum of less accurate to more accurate perception, which indicates that a learner can improve incrementally in perception resulting in incremental improvements in production. In other words, this is a linear relationship between the two scores. If researchers testing the correlation between perception and production are only using a statistical test that will find a linear relationship, non-linear relationships will be masked. Furthermore, this is an inconsistency in the predictions of the SLM that needs to be resolved.
A second point of caution that must be noted about correlation tests is that most analyses represent group scores. Each data point is the combination of one participant’s perception score and production score at a static point in time. As a group, there may be a positive linear relationship between perception and production abilities. But this does not reflect one individual’s change over time. That information could only be ascertained in longitudinal studies that collect information for individual participants over many sessions. This type of study is rare, as it is more difficult to follow participants over the course of years, and few researchers have successfully accomplished documentation over a number of years (e.g., Derwing, Munro, & Thomson, 2008). However, it should be stressed that one-shot testing of a group of participants does not reveal an individual learners’ progression to native-like perception and production. It is important to keep these points about selection of methodology and analyses in mind as I present a sample of correlation studies.

Table 1
Sample of perception and production correlation studies

<table>
<thead>
<tr>
<th>Study</th>
<th>n</th>
<th>L1</th>
<th>L2</th>
<th>Segments</th>
<th>r</th>
<th>p</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baker &amp; Trofimovich (2006)</td>
<td>40</td>
<td>Korean</td>
<td>English</td>
<td>vowels</td>
<td>.73</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Bradlow et al. (1997)*</td>
<td>11</td>
<td>Japanese</td>
<td>English</td>
<td>i, l</td>
<td>.85</td>
<td>.001</td>
<td>pre-train</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.55</td>
<td>.079</td>
<td></td>
</tr>
<tr>
<td>Flege et al. (1999)</td>
<td>72</td>
<td>Italian</td>
<td>English</td>
<td>vowels</td>
<td>.64</td>
<td>significant</td>
<td></td>
</tr>
<tr>
<td>Hattori (2009)</td>
<td>36</td>
<td>Japanese</td>
<td>English</td>
<td>i, l</td>
<td>.53</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Levy &amp; Law (2010)</td>
<td>27</td>
<td>English</td>
<td>French</td>
<td>y</td>
<td>-.15</td>
<td>.274</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>u</td>
<td>-.17</td>
<td>.235</td>
<td></td>
</tr>
<tr>
<td>Peperkamp &amp; Bouchon (2011)</td>
<td>17</td>
<td>French</td>
<td>English</td>
<td>i-i</td>
<td>-.06</td>
<td>&gt;.1</td>
<td></td>
</tr>
<tr>
<td>Schneiderman et al. (1988)</td>
<td>19</td>
<td>various</td>
<td>French</td>
<td>unspecified</td>
<td>.50</td>
<td>&lt;.05</td>
<td></td>
</tr>
<tr>
<td>Sebastian-Gallés &amp; Baus (2005)</td>
<td>27</td>
<td>Catalan</td>
<td>Spanish</td>
<td>e-e</td>
<td>yes</td>
<td></td>
<td>no stats</td>
</tr>
<tr>
<td>Zampini (1998)</td>
<td>13</td>
<td>English</td>
<td>Spanish</td>
<td>p</td>
<td>-.5</td>
<td>unreported</td>
<td>t=1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p</td>
<td>-.19</td>
<td>unreported</td>
<td>t=2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p</td>
<td>.14</td>
<td>unreported</td>
<td>t=3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>b</td>
<td>-.54</td>
<td>unreported</td>
<td>t=1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>b</td>
<td>-.35</td>
<td>unreported</td>
<td>t=2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.17</td>
<td>unreported</td>
<td>t=3</td>
</tr>
</tbody>
</table>

*Although Bradlow et al. (1997) did not report correlation analyses, they provide perception and production scores for all participants, and Pearson correlation analysis was calculated with the data and included in this table.
Table 1 presents a sampling of nine studies that reported correlation analyses of L2 learners’ perception and production scores. As can be seen, the studies are representative of a variety of L2 segments, first and second language combinations, testing methods, and even participant sample sizes. Possibly a result of this diversity, it can also be seen that the outcomes are equally varied. Four of these studies found a strong to moderate correlation between perception and production, two found no statistically significant correlation, and three reported mixed results. With the largest sample size, Flege et al. (1999) tested seventy-two Italian speakers’ ability to perceive and produce English vowels. In a simple correlation, there was a moderate relationship between participants’ discrimination scores and native speaker (NS) identifications of their productions. Flege and his colleagues partialled out L1 use, age of arrival, and length of residence, but the resulting $r$ values were not statistically different than the simple correlation. The results of this study, which is representative of four similar results in Table 1, support the SLM hypothesis that accurate production correlates with accurate perception.

On the other hand, Zampini (1998) tested thirteen NSs of English on their perceptions and productions of the Spanish /p/ and /b/ consonants, specifically investigating the voice onset time (VOT). Although these segments exist in both Spanish and English, their VOTs differ dramatically. As Zampini notes, in English, /p/ has a VOT of 58 ms and /b/ a VOT of 1 ms. On the other hand, Spanish /p/ and /b/ have VOTs of 4 ms and -138 ms, respectively. Zampini measured the Spanish learners’ productions and perceptual boundaries of these segments at the beginning of their foreign language course (time 1), immediately after the students learned about Spanish VOTs (time 2), and at the end of the course, nine weeks after the second session (time 3). This study provides a rare example of longitudinal data of learners’ change over time. At the beginning of the semester, there was actually a negative correlation between the modalities;
many students who were the best producers of the short and negative VOTs had the longest perceptual boundaries. Throughout the semester, this correlation became weaker, and then shifted to a positive correlation by the third session. Zampni concludes that in the case of this contrast, it could be that accurate production must be attained before accurate perception is reached. Short and long lag time for voicing is a feature that already exists in the learners’ L1 phonology. It is possible that the students implicitly learned how to produce Spanish-like VOTs, but the explicit knowledge necessary for perception was delayed. Nevertheless, it was unreported if the \( r \) values even reached significance. Without formal statistical testing, and given the small sample size, this study’s results should be interpreted with caution.

The findings in Zampini (1998), however tentative, suggest it might be possible for learners to produce accurate L2 segments before they perceive them correctly. This goes against the SLM prediction that accurate perception must come first. Because one study alone cannot answer an important theoretical question, it is important to determine if other experiments have found similar results. In fact, this finding had already been noted by two studies, which are summarized in Table 2. Goto (1971) and Sheldon and Strange (1982) tested NSs of Japanese who were considered good and poor producers of English /\( \alpha \)/ and /l/ on their perceptions of the same phonemes. Sheldon and Strange (1982) carefully balanced the phonetic context of the two phonemes (i.e., word initial, word initial cluster, medial, or word final position, and consideration of stress and height of the following vowel) and employed a more consistent perception test than Goto (1971). But in the end, as shown in Table 2, both studies found similar results among the small samples of Japanese speaker participants. There was a relationship between being a better producer and a better perceiver; as production was more accurate, so was perception. But a closer look at the data showed that some participants were able to reach near
perfect production scores as judged by English NSs, yet they were far from perfect perceivers. These participants’ productions were more accurate than their perceptions, again demonstrating that accuracy in the production modality can precede accuracy in the perception modality. These data suggest that there is a connection between the perceptual and productive systems, but the evidence contradicts the SLM prediction that accurate perception must precede accurate production.

The finding that Japanese native speakers’ productive abilities of /ɹ/ and /l/ can precede their perceptual abilities was replicated many years later in a third study by Masuda and Miyashita (2009). This investigation followed three NSs of Japanese over the course of 3.5 years in the United States and collected both perceptual data and productions of the learners at three points in time. All of the learners exhibited better production accuracy than perception accuracy of the /ɹ/-/l/ distinction in at least one of the three sessions. The researchers also noted that the learners’ production accuracy steadily and consistently increased over time, but this was not the case for their perceptual accuracy, which inconsistently improved and declined over time. Although the sample sizes of these three studies of Japanese learners of English (Goto, 1971; Masuda & Miyashita, 2009; Sheldon & Strange, 1982) are all very small and the findings should

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Perception Score</th>
<th>Production Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good Producers (Sheldon &amp; Strange)</td>
<td>5</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Good Producers (Goto)</td>
<td>6</td>
<td>25</td>
<td>6</td>
</tr>
<tr>
<td>Poor Producers (Sheldon &amp; Strange)</td>
<td>1</td>
<td>26</td>
<td>35</td>
</tr>
<tr>
<td>Poor Producers (Goto)</td>
<td>5</td>
<td>36</td>
<td>49</td>
</tr>
</tbody>
</table>

*Note. Perception scores reflect percent of errors when identifying NS of English productions. Production scores reflect percent of errors when NSs of English identified the Japanese speakers’ productions.*
not be generalized across the entire population, the consistency of their results is worth noting. In these studies for some learners, accurate production did precede accurate perception. This finding could be an indication that the SLM postulate cannot be applied across all L2 learning experiences. In some cases, the target phonemes, L1 and L2 combinations, and other substantive features may call for qualifications of this claim.

A limitation of the correlation approach adopted in studies such as those reviewed here is that the type of task researchers use will influence the existence of a correlation. Although Hattori (2009) is listed in Table 1 as showing a strong significant correlation between the two modalities, this researcher ran many correlations between multiple tasks, but only one production measure was found to correlate with perceptual identification scores. In fact, Sebastian-Gallés and Baus (2005) noted that the perception tasks researchers select will reflect different levels of processing, including reliance on acoustic information, working memory, or categorical knowledge, which could ultimately impact correlation outcomes. These researchers tested Catalan-Spanish bilinguals in a categorization task, a gating task, and a lexical decision task, and they found that when the three perception tests were compared against each other, they did not yield the same results. The categorization task was easiest, as 68% of Spanish dominant bilinguals scored at native Catalan levels; the gating task was slightly more difficult with 47% of Spanish bilinguals scoring like Catalan dominant bilinguals; and the lexical decision task was the hardest with only 18% scoring like Catalan dominant participants. Although the tasks seem to show graded levels of difficulty, ten percent of the Spanish dominant participants did not perform better on the purportedly easier tasks and worse on the harder tasks. The results from these two studies indicate that the type of perceptual task used in an analysis will affect the outcome and correlation.
Overall, the sample of studies reviewed in this section suggest two conclusions: (1) there seems to be a moderate correlation between perception and production, but this may be mediated by first and second language combinations, target segments, or tasks and (2) accurate perception does not necessarily have to precede accurate production. Again, it should be emphasized that these are only potential hypotheses because the group of correlation studies that I sampled was not exhaustive. At this point, it is not possible to see patterns that would emerge from a larger and more thorough synthesis. This domain of literature is ready for a meta-analysis in the future, which can make more assertive claims about the correlation between perception and production and how it can be affected by language combinations, target segments, and even tasks.

### 2.2.2. Perception training

Perception training for adults on L2 phonemes is largely successful. Studies have found significant perceptual gains whether the type of training uses natural or synthetic tokens (e.g., Iverson & Evans, 2009; Kondaurova & Francis, 2010), whether the training is held in a classroom, laboratory, or participants’ home (e.g., Aliaga-Garcia & Mora, 2009; Flege, 1995; Iverson, Pinet, & Evans, 2012), or whether it lasts for three months or two minutes (e.g., Gomez Lacabex, García Lecumberri, & Cooke, 2009; Escudero, Benders, & Wanrooij, 2011). This strongly supports the SLM hypothesis that L2 phoneme perception can improve after sufficient exposure to the target phones, and that this capability to create new phonetic categories lasts through adulthood.

If there is a connection between the perceptual and production systems, alteration in one of the modalities may show effects in the other. To investigate this, Bradlow, Pisoni, Akahane-Yamada, and Tohkura (1997) conducted a perceptual training program with NSs of Japanese on the English segments /ɹ/ and /l/ and also tested for gains in production. Eleven subjects
underwent forty-five sessions of high variability identification training over the course of three to four weeks. Participants’ perception improved from 65% to 81% accuracy, and their productions showed improvements on two different NS rating measures. Over 450 published articles have cited this study as evidence that perceptual training alone can lead to production gains, or in more general terms, that there is a connection between the two modalities. However, the field should be cautious of placing such a high emphasis on one study to prove such an important theoretical claim. First of all, this study targets only one set of languages and one segmental contrast. Additionally, the experimental group sample size was very small, only eleven. The findings of Bradlow et al. (1997) cannot be generalized and should not be the only piece of evidence used to suggest a connection between the perception and production systems.

As a response to this overreliance on one article, a meta-analysis on the relationship between L2 perceptual training and production gains was conducted by Sakai and Moorman (submitted). Although hundreds of potential studies were identified, thirty studies passed all inclusion and exclusion criteria to be synthesized. Of those thirty, only eighteen included sufficient data for an effect size to be calculated. One of the reasons many experiments were excluded at this very last stage was a result of perceptual training programs including production practice: Fifty-three studies were eliminated from the final pool because they conflated perception training with production practice. As Sakai and Moorman stressed, if researchers are interested in the question of cross-modal training benefits, that is, whether gains can spread across the modalities, then it is imperative that the modalities be kept entirely separated. Of course, many of these experiments were based in realistic, real-world settings in which listening and speaking do naturally co-occur. However, if the intention is to test theoretical hypotheses, strict separation should be a priority.
Table 3

Within-group Cohen's d effect size scores for eighteen studies, with means (SD) presented in the last row of each column (from Sakai & Moorman, submitted)

<table>
<thead>
<tr>
<th>Study</th>
<th>Independent Experimental Groups</th>
<th>Perception Effect Size</th>
<th>Production Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anderson (2011)</td>
<td></td>
<td>.44</td>
<td>.31</td>
</tr>
<tr>
<td>Bradlow et al. (1997)</td>
<td></td>
<td>1.38</td>
<td>.40</td>
</tr>
<tr>
<td>Counselman (2010)</td>
<td></td>
<td>-</td>
<td>.63</td>
</tr>
<tr>
<td>Gomez Lacabex &amp; Garcia Lecumberri (2010)</td>
<td></td>
<td>-</td>
<td>.75</td>
</tr>
<tr>
<td>Han (2002)</td>
<td></td>
<td>-</td>
<td>.88</td>
</tr>
<tr>
<td>Hazan et al. (2005)</td>
<td>Audio Only</td>
<td>1.45</td>
<td>.09</td>
</tr>
<tr>
<td></td>
<td>Audio-Visual (Natural)</td>
<td>1.11</td>
<td>.42</td>
</tr>
<tr>
<td></td>
<td>Audio-Visual (Synthetic)</td>
<td>1.15</td>
<td>.16</td>
</tr>
<tr>
<td>Herd et al. (2013)</td>
<td></td>
<td>.51</td>
<td>.49</td>
</tr>
<tr>
<td>Huensch (2013)</td>
<td></td>
<td>1.03</td>
<td>1.31</td>
</tr>
<tr>
<td>Lambacher et al. (2005)</td>
<td></td>
<td>.47</td>
<td>.43</td>
</tr>
<tr>
<td>Lengeris (2009)</td>
<td></td>
<td>-.04</td>
<td>-.08</td>
</tr>
<tr>
<td>Motohashi (2007)</td>
<td>Audio Only</td>
<td>.73</td>
<td>1.31</td>
</tr>
<tr>
<td></td>
<td>Audio-Visual</td>
<td>1.87</td>
<td>1.75</td>
</tr>
<tr>
<td>Nobre-Oliveira (2007)</td>
<td>Natural Tokens</td>
<td>1.34</td>
<td>-.03</td>
</tr>
<tr>
<td></td>
<td>Synthesized Tokens</td>
<td>2.19</td>
<td>1.34</td>
</tr>
<tr>
<td>Reis &amp; Nobre-Oliveira (2007)</td>
<td></td>
<td>-2.14</td>
<td>.84</td>
</tr>
<tr>
<td>Soler-Urzua (2011)</td>
<td>Text-to-Speech</td>
<td>.51</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td>Non Text-to-Speech</td>
<td>-.10</td>
<td>.04</td>
</tr>
<tr>
<td>Thomson (2007)</td>
<td>Long Vowel Training</td>
<td>.70</td>
<td>.54</td>
</tr>
<tr>
<td></td>
<td>Select Vowel Training</td>
<td>1.36</td>
<td>.55</td>
</tr>
<tr>
<td>Underbakke (1993)</td>
<td></td>
<td>1.40</td>
<td>.33</td>
</tr>
<tr>
<td>Wang (2002)</td>
<td></td>
<td>2.84</td>
<td>.75</td>
</tr>
<tr>
<td>Yeon (2004)</td>
<td></td>
<td>2.24</td>
<td>.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.97 (1.03)</td>
<td>0.58 (0.47)</td>
</tr>
</tbody>
</table>

Note. Counselman (2010), Gomez Lacabex & Garcia Lecumberri (2010), and Han (2002) did not administer a pre- and post-test of perception.

What then did the findings of the meta-analysis show about the possibility of perception-to-production cross-modal gains, after stringently controlling for perception-only trainings? Of the thirty studies included in the synthesis portion of the meta-analysis, 73% found crossover effects into some sort of gain on at least one production test. It is important to note that many of these studies did not report gains on all of the segments that were trained (e.g., Anderson, 2011; Herd, 2011; Nobre-Oliveira, 2007; Reis & Nobre-Oliveira, 2007) nor on all tasks used to test production (e.g., Soler-Urzua, 2011). Table 3 is reproduced from the meta-analysis with
permission, and it details the within-group effect sizes of both the perception and production modalities for the eighteen investigations that passed all inclusion and exclusion criteria.

Because each experiment used a variety of tasks and tests, effect size calculation is important to equalize the dispersion of measurements and scoring systems. From the eighteen studies, twenty-four effect sizes were calculated; these data measure the magnitude of the effect perception training had on production outcomes. The range of effect sizes was -.08 to 1.75, with an average effect size of 0.58. According to the standards suggested by Plonsky and Oswald (2014) for the interpretation of Cohen’s $d$ effect sizes in second language acquisition, this average shows that there is a small-sized effect of perception training on production.

Figure 1 presents the effect sizes grouped by the target segments that were trained. The bar graph shows how perception trainings that targeted vowels and sonorants had smaller effects on production outcomes than the obstruent category. This difference could be due to place of articulation and ease of deciphering vocal tract articulations through the perceptual mode. This may be evidence of phoneme salience depending on the interpretability of the physical motor origins of sounds. If there is a continuum of more detectable articulations to less detectable, then those phonemes which have more easily detected articulations might be more susceptible to cross-over effects from perception to production. For example, the study that showed the highest effect size for production gains was a dissertation by Motohashi (2007). The target segments of this study were Japanese geminates. Perhaps it is the case that the motor control of length distinction in consonants is easily detectable in the acoustic information, and is also coupled with very easy motor demands, ultimately prompting a strong outcome in production gains. A study by Hisagi and Strange (2011) supports this speculation, as findings showed that simply drawing American English listeners’ attention to the Japanese length distinction was sufficient assistance
for them to discriminate the contrasts well above chance (Hisagi & Strange, 2011). On the other end of the extreme, vowels involve slight movements of the tongue body, which may be less easily interpreted from the acoustic information, therefore causing a smaller effect in the production changes.

Sakai and Moorman’s (submitted) meta-analysis offers insights from a broader perspective about the effects of perceptual training on production, by revealing that the effects of perception training on production gains are moderate. It seems that perceptual training induces some gains in productive knowledge, at least for some segments and on some tasks. As the perceptual L2 system sharpens the understanding of a phoneme and adds information about how to produce the phone with more target-like accuracy, the productive system can use that same information to better articulate the phone as well. Learners either implicitly or consciously use that information in order to better produce the sounds.
2.2.3. Production training

I have argued in the preceding section that perception training programs seem to have a strong to weak effect on production gains depending on the target phoneme, which nonetheless indicates that the productive system is accessed through the perceptual modality. When considering training studies, it is also possible to consider the reverse direction: Will production training programs affect the perceptual system? There is quite a large body of SLA literature that investigates pronunciation training and instruction. In fact, a sufficient number existed for Lee, Jang, and Plonsky (2015) to conduct a meta-analysis with a total of eighty-six experimental or quasi-experimental reports. For experimental groups that underwent pronunciation instruction, the within-group Cohen’s $d$ effect size was $d=.89$, which means that instruction generally has a substantial impact on learners’ production. The findings will be further reviewed later in the Results and Discussion Chapters. For now, however, it must be noted that in the normal communicative setting, language learners will listen to speech as they practice pronunciation with more target-like accuracy. Thus, pronunciation training in traditional environments cannot aid in establishing a bidirectional relationship between perception and production. In theoretical terms, on the other hand, production training should be isolated from perception, which must mean no aural input is provided to participants. This is not a particularly intuitive design, which perhaps helps to explain why not many researchers have attempted or even considered training production in this way.

Most pronunciation training studies provide participants with an auditory example to be analyzed or imitated. For example, Hattori (2009) conducted a study of Japanese learners targeting the English /u/-/l/ distinction. Participants were trained in one-on-one articulatory sessions with real-time spectrograms, and were tested in pre- and post-tests with both perception
and production tasks. The researcher sought to determine if production training could influence perception scores, but throughout the training, participants heard multiple tokens of the target segments. They listened to their own recordings that were spectrally altered to be more target-like, they watched a video recording of a model pronouncing the segments, and they listened to and repeated after their instructor. Although the training did focus on the articulatory aspects of English /u/ and /l/, the participants repeatedly heard many examples of the segments throughout the training. This study, and others like it, cannot claim that articulatory training leads to perception gains if hearing the auditory stimuli might have actually caused alterations in the perceptual modality. There are many other pronunciation or articulatory training programs that cannot truly reveal a connection to the perceptual modality because the researchers did not control for repeated and dense exposure to the target sounds (e.g., Gomez Lacabex, García Lecumberri, & Cooke, 2008, 2009; Hirata, 2004; Kissling, 2015; Perez-Gamboa, 1999). Only more recently, Hattori and Iverson (2010) called for production training studies to be implemented without listening tasks if researchers are genuinely interested in understanding the theoretical connection between the modalities.

After a thorough search of the literature, I uncovered five studies that may have attempted production training in the absence of auditory input. However, due to a lack of detailed reporting in methodology sections, it is unclear if some of them actually did allow participants to hear audio samples. I review them below.

Flege (1988) trained a Spanish-speaking participant to produce English vowels using a glossometer, which provided visual, real-time feedback of the tongue shape. Remarkably, the participant was able to pronounce the /i/-/ı/ contrast correctly after only ten minutes of training. The methodology does not include any mention as to whether the participant did or did not hear
samples to imitate. Additionally, Flege did not test for improvements in the perceptual modality either. This is also the case for a study conducted by Schmidt and Beamer (1998), in which the researchers used electropalatography to train three Thai speakers on six English consonants. The report does not include any statement as to whether the participants heard audio samples, nor did the researchers conduct perception pre- and post-tests.

The three remaining studies are considerably more recent and either explicitly indicated their participants were not allowed to hear samples of the target segments (i.e., Herd, 2011; Warsi, 2001) or the authors reported the methods with such detail that it is unlikely the training involved audio samples (i.e., Liakin, Cardoso, & Liakina, 2013). Moreover, all three studies conducted perception pre- and post-tests.

The earliest of the studies is Warsi (2001). In this experiment, twelve NSs of Japanese were trained on the English /ɹ/-/l/ distinction in eight one-hour sessions over the course of four weeks. Each session began with a presentation of diagrams and descriptions of the placement of articulators during both segments. The researcher emphasized that the instructor did not model the target phones; instead, he provided immediate feedback of the participants’ productions. Nevertheless, participants were in the same room together during the training sessions, and presumably heard each other model the sounds. An English NS judged participants’ productions for the pre- and post-test, and overall, participants showed an improvement in production of /ɹ/ and /l/ in almost all contexts at word and sentence level utterances. For perception, participants completed an identification task in which target sounds were embedded in sentences. For example, after hearing a sentence, participants filled in the missing word, one of a minimal pair (e.g., collect or correct). In this task, participants improved identification of /l/ in consonant clusters and /ɹ/ in medial and consonant cluster positions only. The researcher concludes by
saying that accurate production can precede perception, a point that was noted in the previous section with this same population and target phones. However, the control group also showed significant gains on the perception test. This is not surprising because all participants were enrolled in English as a second language courses, and extra-experimental learning is likely in this environment. Therefore, the results of this study are inconclusive if articulatory production training can lead to perceptual gains.

Next, Herd (2011) trained ten NSs of English in Spanish /ɾ/, /r/, and /ð/. Participants analyzed and compared their own productions to the waveforms and spectrograms of six NSs of Spanish in six training sessions, over the course of two to three weeks. The participants never heard the audio of the NS samples throughout the training. Results from production pre- and post-tests showed that participants improved in their productions of /r/ only. In a perception identification task, pre- to post-test scores showed that they did not improve on the contrasts overall, but when analyzed individually, it was revealed that they significantly improved on identifying the /ɾ/-/r/ contrast. It should be noted that Herd (2011) is also included in the perception training meta-analysis of the previous section and in Table 3. Participants who received perceptual training showed a strong effect in the production modality (d=1.55). This study isolated the modalities very cleanly, and was able to show a strong crossover effect in one direction (perception to production), but a much weaker effect in the other (production to perception). This could be due to the lack of gains in the production training itself.

Finally, Liakin, Cardoso, and Liakina (2013) trained native or near-native speakers of English in the French production of /y/. Participants were all beginner-level French learners who had not yet acquired the French vowel. The experimental group used a free application on electronic mobile devices that utilized automatic speech recognition. Fourteen participants
completed five twenty-minute sessions over five weeks. They spoke a word into the mobile
device and received immediate written feedback on the accuracy of their productions. For pre-
and post-tests, NSs of French marked participants’ productions of the vowel as correct or
incorrect. The productions of the experimental group improved significantly, but their perception
scores on an identification task showed that this production gain had no influence on their
perceptions of the vowel.

What can be concluded from these rare studies that attempted to control for aural input in
order to test for effects of production training on the perception modality? In general, they were
not able to find gains (with the exception of one contrast in Herd, 2011). In this small sample of
studies, it seems that improvements in production do not have a strong effect on perception,
which means the connection between the modalities may not be bidirectional. At least in these
few available studies, as learners are sharpening the motor skills needed to produce an accurate
sound, they are not getting better at creating a category for that sound. One explanation could be
an effect due to variability. In perception training studies, one feature of the training that
consistently produces positive results is high variability in the auditory exposures during training
(Iverson, Hazan, & Bannister, 2005). High variability phonetic training (HVPT) offers listeners a
chance to hear many tokens of a phonetic category, from different speakers, in different phonetic
contexts, and with various other features (e.g., duration). In production training, by necessity,
there is no variability in the number of speakers, because the participant can only produce one
voice. But production training programs might be able to offer practice with a variety of phonetic
contexts and other features. At this point, the number of studies that attempt to control for
auditory influence in production training is too small to be able to generalize any findings.
An issue that has not been addressed by any previous study on isolated production training is the possibility that the existence of self-produced auditory exposure might be a perception-driven component responsible in part for any observed perceptual benefits of the production training. All of the studies mentioned in this section allowed learners to hear their own voices, thus they were not isolated from perceptual input. Even if a training program did not provide models of target segments, a person’s own voice will provide aural input. Self-produced input reaches the speaker’s auditory canal through two mediums. First, and most obviously, sound waves travel through the air to reach a speaker’s ears. Self-provided input can be controlled with noise-cancelling headphones. But this still leaves one other mode through which a person will hear his or her own voice: Sound waves also travel internally through the head. This is called bone-conducted sound (Stenfelt & Goode, 2005). Through this medium, it is possible to hear every spoken word, albeit in a damped form. In order to be extremely cautious of eliminating all aural input in a production-only training design, the participant’s voice that travels through air and bone conduction must be cancelled out. To the best of my knowledge, no production training study has ever attempted to test production (or perception) benefits from a training regimen that carefully eliminates the possibility of auditory access to self-produced models. The present dissertation will attempt to do so in Experiment 2 for the first time.

This dissertation comprises two experiments that attempt to fill in a gap found in the production-only training of L2 phonemes, by training native speakers of Spanish in a difficult-to-perceive English vowel contrast. In the remainder of this chapter, I will review the relevant literature that helps provide background information on the L2 phonological development for this population of learners.
2.3. Second language vowel development of native speakers of Spanish learning English

Native speakers of Spanish who are studying English as a second language are a well-investigated population in the L2 phonology literature. Many studies report phonetic segments that are typically troublesome for them, and this population is well represented in the perception training literature as well. The present dissertation will address the theoretical question of bidirectional cross-modal training effects for perception and production by training L1 Spanish speakers in one vowel contrast, the English /i/ and /ɪ/ vowels. Therefore, in this last section of Chapter 3 I review relevant findings regarding the acquisition of English vowels by L1 Spanish speakers.

The Spanish vowel inventory contains only five phonemes, /i/, /e/, /a/, /o/, /u/ (Cenoz & García Lecumberri, 1999). As native speakers of Spanish begin to learn English, a language that has over twelve contrastive vowels, they face the challenge of hearing the second language vowels through the Spanish system. For native Spanish and Basque speakers, Cenoz and García Lecumberri (1999) determined that Received Pronunciation (RP) simple vowels /i:/, /æ/, /ɒ/, /ɜː/, /ʊ/ were hardest for them to discriminate. In an earlier study, the same authors reported a similar set of RP British English vowels, /i:/, /ɒ/, /ɜː:/, /u:/, as being the most difficult for 117 native Spanish speakers (García Lecumberri & Cenoz, 1997). Iverson and Evans (2009) report that /i/-/ɪ/, /a/-/ʌ/-/ɑ/, and /ɒ/-/ɔ/ are the most often confused British English vowel pairs for Spanish speakers.\(^1\) Morrison (2008), McCracklin (2012), and Garcia (2014) have also studied the difficulties of NSs of Spanish to acquire the Canadian and American English /i/-/ɪ/ and /ɑ/-/ʌ/ contrast.

\(^1\) It should be noted that in all of these studies, the target forms were extracted from the British vowel inventory, which is not the same as the General American English vowel system.
One feature present in the above lists of phonemes is the tense-lax distinction. Many perceptual training studies have reaffirmed that the tense-lax distinction is a difficult contrast for English learners across a variety of L1 backgrounds; Kondaurova and Francis (2010) and Ylinen et al. (2009) have both focused on the /i/-/ɪ/ pair as the target form in their perceptual training studies. These English vowels can be distinguished by both spectral and durational cues, although the latter is not always a reliable differentiating feature. NSs of English rely mainly on the spectral cues provided by the first three formants (Kondaurova & Francis, 2010); on the other hand, Spanish speakers learning English misdirect attention to vowel duration, while not attending to the differences in spectral cues. Spanish speakers learning Dutch also showed this same reliance on durational cues while attempting to distinguish /ɑ/ and /a:/ (Escudero, Benders, & Wanrooij, 2011). Because Spanish does not contrast vowels based on duration, Kondaurova and Francis (2010) infer that NSs of Spanish have misidentified this cue to be reliable evidence to distinguish between tense and lax vowels.

The difficulties that NSs of Spanish have learning the English vowels have presented a very appealing environment for researchers interested in the plasticity of an adult learner’s native phonetic representations. Motivated by the SLM hypothesis that with sufficient input learners can improve perceptions of L2 sounds, Aliaga-Garcia and Mora (2009), Cenoz and García Lecumberri (1999), Iverson and Evans (2009), Kondaurova and Francis (2010), Gomez Lacabex, García Lecumberri, and Cooke (2009) have all explored the effects of distinct perceptual training on Spanish speakers’ acquisition of English vowels. All of the studies successfully trained NSs of Spanish in the perception of the target vowels.

Possibly the most precise study was conducted by Kondaurova and Francis (2010). They addressed the issue of reliance on durational cues by placing participants in one of three
perceptual training groups: inhibition training \((n=9)\), adaptive training \((n=10)\), and natural correlation training \((n=10)\). All groups underwent four thirty-minute training sessions using an identification task, but the stimuli used during training differed for each group. The inhibition training group heard stimuli that provided high variability in duration in order to deter listeners from using this cue to identify the vowels. The adaptive training group first heard stimuli that were spectrally farther apart, representing extreme values of the tense-lax continuum. Once they achieved more than 80% accuracy identifying these tokens, they were moved to a stimuli set that was spectrally closer on the continuum. Finally, the natural correlation training group was presented with tokens mimicking HVPT training; in this case variability meant spectral and duration differences, rather than number of speakers. For pre- and post-testing, all groups completed perception identification and discrimination tasks. The results showed that all groups improved performance in correct responses and in reaction times. The inhibition training group reached 99% accuracy, the adaptive training group reached 90% accuracy, and the natural correlation training group reached 99% accuracy on the perceptual identification task. Additionally, all three training groups showed generalization of improvement on natural stimuli not used in the training. However, in terms of spectral or duration cue reliance, only the inhibition training group was successful in shifting attention to spectral cues. When compared to a NS English baseline group, the performance in spectral to duration cue reliance was not significantly different. The other two groups, adaptive training and natural correlation, did not achieve this shift in cues.
Chapter 3: Interdisciplinary Contributions to the Perception-Production Conversation

In this chapter, I will introduce information from disciplines that are not typically inspected in the second language acquisition literature or the literature on the development of L2 sound systems. A wide range of scientists from the two disciplines examined in this chapter investigate human perception and production, and knowledge of these other research domains adds value to the overall understanding of the interrelated nature of the two modalities.

3.1. Deaf and hard-of-hearing populations

One area of research that adds insight to the perception-production connection is the literature on the deaf and hard-of-hearing population. Children who are raised in a linguistic environment in which their auditory information is limited can reveal patterns of the association between the speech perception and production modalities, and many of the studies in this domain have asked similar questions as the SLA literature, particularly the correlation between the two modalities and the effects of training programs. In this section, I build on landmark studies in this field, specifically studies conducted by Donna Geffner and David Ertmer, which are little known or utilized in SLA (e.g., see Fey et al., 2011; Ertmer et al., 2013).

In a 1987 monograph, Geffner and Levitt conducted a number of assessments on sixty-seven young children with profound hearing loss. The children were six years old, with an average hearing loss of 104 dB. The average age when hearing impairments were first detected was 1.6 years old, and the average age that the children were fitted with hearing aids was 2.7 years old. This means they were categorized as prelingual when their hearing loss was diagnosed.

---

2 Normal talking is 45-60 dB and a loud concert is 110 dB. If a child has 104 dB hearing loss, it means that 104 dB is the softest sound that can be heard at least 50% of the time.
Geffner and Levitt report earlier studies that consistently find that this population has difficulty producing palatal and alveolar fricatives, affricates, and velar nasals. For vowels, the most common errors are neutralization, diphthongization, and nasalization. In their own tests, a syllable imitation and a picture description task, the researchers found that the children produced vowels with quite high accuracy (77.2%) and consonants with a much lower accuracy (23.9%). For consonant sounds, as the place of articulation moved further back into the vocal tract, the children had a harder time producing them with accuracy: linguadental (22%), alveolar (19%), and velar (4%). Affricates /dʒ/ and /tʃ/ and fricatives /z/, /ʃ/, and /s/ were among the hardest segments for them to produce. This suggests that the participants began to rely on the visual system in order to mitigate the absence of clear acoustic information; phonemes with highly visible articulators were easier to imitate and therefore easier to acquire.

An additional finding that Levitt (1987) reports in a different chapter of the same volume is that “intelligibility of a child’s spontaneous speech production correlates moderately well with his or her hearing level, but shows a higher correlation with measures of speech reception” of both prosodic and segmental features (p. 124). This means that there was a negative correlation between speech intelligibility and degree of hearing loss. Children with 50-70 dB hearing loss scored consistently high on intelligibility ratings, in some cases near perfect scores, but intelligibility of productions steadily decreased as hearing loss increased. This correlation shows that for first language acquisition⁴, there is a strong connection between perception and production abilities. This supports the idea that accurate production is dependent on perception.

---

⁴ For some of the children, the acquisition of spoken language may have been their second language, as sign language was their first language.
A more recent longitudinal study by Ertmer, Jung, and Kloiber (2013) again depicts the necessarily entwined nature of perceptive and productive abilities. Without rich aural exposure to language, children’s development of speech-like production is dramatically hindered. However, when these same children are exposed to a sufficiently robust hearing environment through cochlear implants, their productive abilities rapidly develop. Ertmer, Jung, and Kloiber recorded two groups of children at three-month intervals for each child’s first twenty-four months of robust hearing experience. The first group consisted of twelve children who were diagnosed with profound, bilateral hearing loss. Recording for these children began when they received a cochlear implant, which occurred at various ages from nine to thirty-six months of age (M=21 months). This group, the cochlear implant (CI) group, was compared to a second group of eleven children with normal hearing, named the typically developing group (TD). Their recording sessions began at six months of age. All of the children’s utterances were categorized according to a scale of three developmental stages: precanonical vocalizations (e.g., vowels, squeals, trills), basic canonical syllables (e.g., CV, CVCV), and advance forms (e.g., VC, CVC, diphthongs). Results showed that the CI group development occurred at a more rapid rate than the TD group. After eighteen months of robust hearing, 100% of the CI group had reached speech-like expressions in 60% or more of their utterances. For the TD group, this occurred at twenty-four months of hearing experience. The faster rate of development for the CI group can be attributed to neurological, cognitive, social, and motor development advantages due to the groups’ age difference. These findings apply to the general question of the perception-production connection because they demonstrate that auditory information is absolutely necessary for productive development to occur in the first language.
Although the information about correlations and longitudinal development shows the interconnected nature of perceptive and productive abilities in children with profound hearing loss, there is also evidence from this same population that pronunciation training programs in the absence of full auditory information can be remarkably successful. Ertmer, Stark, and Karlan (1996) trained two nine-year-old children in vowel productions: the first participant had 113 dB hearing loss in his right ear and did not respond to any sounds in his left ear; the second participant could not hear above 500 Hz in either ear, and had 80 to 115 dB hearing loss in both ears. Both participants were selected for the study because they had less than 10% accuracy producing vowels. Participants underwent individual training sessions of thirty minutes, three times a week for five months. The training program utilized real-time visual feedback with spectrograms, and consisted of three stages: spectrogram and formant instruction, information about the articulators, and time to practice producing vowels with the spectrogram and self-judgment of the correctness. At the end of the five months, the first participant significantly improved on all three trained vowels, /i/, /æ/, and /o/, and reached 90% accuracy on these vowels after only seven, five, and four training sessions, respectively. The second participant was not as successful, but also improved vowel productions. This participant reached 100% accuracy producing /a/ after fifteen training sessions and 100% accuracy producing /æ/ after six sessions, but she was not able to reach 70% on her third vowel, /i/, before the end of the study. Although the sample size of this study is admittedly small, this type of successful training is not uncommon in this literature (e.g., Bacsfalvi, Bernhardt, & Gick, 2007; Ertmer & Maki, 2000; Gallagher, 2013; Massaro & Light, 2004; Osberger, 1987). Production can improve in the absence of accurate perception. However, none of these training studies fully eliminated acoustic
cues nor did they test the participants’ perceptual gains, so it is unknown if participant perceptions also improved with their productions.

If articulatory training with feedback is sufficient to improve pronunciation to be more target-like for individuals with limited to no perceptual abilities, accurate perception may not be necessary for NNSs to consistently produce a recognizable and intelligible version of a phoneme. In sum, the studies with deaf and hard-of-hearing populations reviewed in this section both support a connection between the modalities and disconnect them as well.

3.2. Neurocognition

A second interdisciplinary field that can provide valuable information about the overlap in perception and production processing is neurology. With advances in technology and equipment, neuroscientists are able to detect which parts of the brain contribute to various functions, including speech communication. A review of the first twenty years of PET and fMRI scans in speech research was published by Price in 2012. This review compiles information from hundreds of reports on heard, spoken, and read speech. I have excerpted a portion of a table printed in this study in Table 4, which depicts the most current understanding of the parts of the brain activated by heard and spoken speech at the time of the article’s publication.

The overview of parts of the brain associated with the auditory and articulatory processes in Table 4 represents the best knowledge based on the literature from 2007 to 2011. As I will discuss below, it does not incorporate a growing body of literature that presents compelling evidence of motoric contributions to auditory processing. This classical view of perceiving and producing speech separates the systems as two independent structures. Hearing speech predominantly activates the superior temporal gyrus, and producing speech activates the motor
Table 4

Regions of the brain associated with perceiving and producing speech (summary of a portion of Table 2 in Price, 2012)

<table>
<thead>
<tr>
<th>Process</th>
<th>Brain Region</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Auditory processing of speech and non-speech sounds</strong></td>
<td></td>
</tr>
<tr>
<td>Acoustic processing</td>
<td>L and R superior temporal gyrus</td>
</tr>
<tr>
<td>Rapid transitions</td>
<td>L posterior superior temporal gyrus/planum temporale</td>
</tr>
<tr>
<td>Acoustic complexity</td>
<td>L anterior superior temporal gyrus</td>
</tr>
<tr>
<td><strong>Speech selective auditory processing (phonological processing)</strong></td>
<td></td>
</tr>
<tr>
<td>Articulatory recoding</td>
<td>L ventral pars opercularis</td>
</tr>
<tr>
<td><strong>Covert articulatory planning for production of speech sounds</strong></td>
<td></td>
</tr>
<tr>
<td>(phonological output)</td>
<td></td>
</tr>
<tr>
<td>General action selection</td>
<td>L and R premotor cortex, L dorsal premotor cortex</td>
</tr>
<tr>
<td>Sequencing motor plans</td>
<td>Supplementary motor cortex, pre-supplementary motor cortex,</td>
</tr>
<tr>
<td></td>
<td>L dorsal pars opercularis</td>
</tr>
<tr>
<td>Orofacial motor planning</td>
<td>L ventral pars opercularis, L ventral premotor cortex</td>
</tr>
<tr>
<td>Auditory expectation</td>
<td>Planum temporale, ventral supramarginal gyrus</td>
</tr>
<tr>
<td><strong>Overt articulation</strong></td>
<td></td>
</tr>
<tr>
<td>Motor execution</td>
<td>Supplementary motor cortex, p zone</td>
</tr>
<tr>
<td>Orofacial motor activity</td>
<td>L and R precentral and postcentral (rolandic cortex)</td>
</tr>
<tr>
<td>Timing of motor output</td>
<td>Cerebellum (IV &amp; V), paravermal, RCB</td>
</tr>
<tr>
<td>Breathing control</td>
<td>Ventral and lateral thalamus, L anterior insula</td>
</tr>
</tbody>
</table>

*Note. “L” signifies left, and “R” is right.*

cortex. The only parts of the brain that are represented in both systems are the planum temporale and the pars opercularis. The planum temporale is located on the dorsal surface of the posterior temporal gyrus, and the pars opercularis is part of the region known as Broca’s area. From this perspective, it appears that perceiving and producing speech are generally separate processes.

Even though Price’s (2012) review of the neurolinguistic literature includes recent studies, she did not address the growing body of literature that presents persuasive evidence of motoric contributions to auditory processing. Over the past decade, many neuroimaging studies have begun to find overlap of the motor cortex in passive speech listening. One study, conducted by Wilson, Saygin, Sereno, and Iacoboni (2004), used fMRI scanning on ten participants as they listened to and later repeated monosyllabic nonwords. During the perception portion of the task,
all ten participants showed activation in the ventral premotor cortex spanning to the primary motor cortex. Four participants showed bilateral activation, two in the left-hemisphere, and four in the right hemisphere. Brain regions that were activated during the production task overlapped with the perception task by 73%. The researchers conclude, “These findings are consistent with the view that speech perception involves the motor system in a process of auditory-to-articulatory mapping across a phonetic code with motor properties” (p. 702). In the years following this study, many others have corroborated the findings that the motor cortex is involved in passive speech listening and singing (e.g. Chen, Penhune, & Zatorre, 2008; Pulvermuller, Huss, Kherif, Moscoso del Prado Martin, Hauk, & Shtyrov, 2006; Skipper, Nusbaum, & Small, 2005).

One shortcoming of fMRI scanning is that this methodology measures hemodynamic response activation, which is the amount of blood flow to particular regions of the brain, in response to different stimuli or tasks. This means that fMRI results only report correlations between blood flow and processing of certain tasks (D’Ausilio, Pulvermuller, Salmas, Bufalari, Begliomini, & Fadiga, 2009), and different methods are necessary to determine that specific regions of the brain actively contribute to specific tasks or processing. In the case of speech perception, researchers have begun to use transcranial magnetic stimulation (TMS) to temporarily stimulate or disrupt activation of a specific region of the brain to measure effects on behavioral task accuracy or reaction times. D’Ausilio and his colleagues utilized TMS pulses to stimulate the lip and tongue representations in the premotor cortex before participants completed a phoneme discrimination task involving [p], [b], [t], and [d] syllables masked in white noise. When TMS pulses stimulated the lip area, participants were faster and more accurate at responding to [p] and [b] trials, but not [t] and [d] trials. The reverse was also true: When the
tongue areas were stimulated, participants were faster and more accurate discriminating the coronal sounds, but not the bilabials. This study provides strong evidence that the motor regions of the brain contribute to speech perception.

Further investigations suggest that the motor cortex is indeed activated in passive speech perception, but it only contributes to speech perception in difficult-to-perceive conditions. When task demands are increased in various ways, auditory-to-articulatory mapping can provide additive information to the compromised acoustic signal. D’Ausilio et al. (2009) asked participants to discriminate sounds masked in white noise, which increases the difficulty of the task. In this compromised listening environment, motor contributions were found. In a subsequent study D’Ausilio, Bufalari, Salmas, and Fadiga (2012) added an experimental group to the previous TMS study. Participants again received the stimulating TMS pulses in the lip and tongue regions of the premotor cortex, but this time, discrimination stimuli were not masked in white noise. In this ideal listening condition, participants’ accuracy and reaction times were not affected by TMS. It is possible that the tasks were so simple that scores reached ceiling. The researchers concluded that the “motor system may play a more important role when sensory information is incomplete, eventually by filling in the gaps via attentional or top-down processing” (p. 886), but not in an ideal listening environment. Moreover, Hickok, Holt, and Lotto (2009) suggest that motor contributions are not necessary for speech perception, but they can account for approximately 10% of performance on perception in compromised contexts. Compromised listening environments that show activation in the motor cortex include listening to non-native phonemes (Callan, Tajima, Callan, Kubo, Masaki, & Akahane-Yamada, 2003; Wilson & Iacoboni, 2006) and foreign accented, native language phonemes (Callan, Callan, & Jones, 2014). In ideal listening conditions, motor contributions may not be necessary to perform
discrimination tasks. However, in life outside of a laboratory, listening conditions are often far from ideal, which could indicate that the auditory-to-articulatory mapping is often utilized.

Neurolinguistic studies provide evidence of motor contribution to speech perception, and there is also evidence that supports contributions in the opposite direction. Auditory areas of the brain are recruited in speech production. Covert speech production involves moving the articulators to create silent speech. Numerous studies have found auditory cortex activation in these conditions (e.g., Parker Jones, Seghier, Kawabata Duncan, Leff, Green, & Price, 2013; Price, Crinion, & MacSweeney, 2011; Sato, Troille, Menard, Cathiard, & Gracco, 2013). Price and her colleagues (2011) asked participants to produce covert speech and non-speech mouth movements in an fMRI scanner. Covert speech, but not non-speech mouth movements, activated Broca’s area, Wernicke’s area, and the left posterior superior temporal sulcus. The researchers believe that “prior experience producing speech sounds leads to the automatic and covert generation of auditory and phonological associations” (p. 1). In regular speech, the auditory cortex could be used to produce an internal model, which can be compared against the actual output. In other words, the auditory activation could be predictive and used to modulate one’s own productions. If the produced stream does not match the acoustic prediction, a speaker can adjust the vocal tract to create a sound that was intended.

The overlap of the perceptive and productive systems in terms of activated brain regions seem to have evolved to make speech processing more efficient, both for the perceiver and the producer. Overall, many of the studies cited in this section suggest that both speech perception and production are a dual-integrated system. Perception is not purely acoustic, and production is not purely motoric, but both are a sensorimotor experience.
Chapter 4: Research Questions

I hope to have shown in Chapters 2 and 3 that the literature from SLA, the deaf and hard-of-hearing population, and neurocognition provide a broad scope of information about how the perceptive and productive systems are intertwined and how they are separate. In Chapter 2, I argued that second language correlation analyses reveal there may be a moderate relationship between group scores on perceptive and productive tasks at one point in time, and some individuals show better production skills than perceptive skills. Moreover, these findings may be moderated by L1-L2 combination, target segments, and tasks. Training studies reveal that access to one modality through the other may not be bidirectional. Perception training moderately improves production, again depending on the target segment and task, but the few production training studies that exist do not show effects in the perception modality. More research is needed in this area of strictly isolated production training. The research on the deaf and hard-of-hearing population reviewed in Chapter 3 demonstrates that rich auditory information is necessary for production development, but production can also improve independent of complete access to aural input. Finally, in neurolinguistics, the processing areas have classically been associated with separate regions of the brain, but more recently, researchers are finding that motor regions of the brain may provide auditory-to-articulatory information in hard to hear situations. Additionally, the auditory cortex may provide information to regulate the alignment of intended and actually produced speech.

The present dissertation study addresses these issues in the extant literatures reviewed. It is a first attempt to strictly disassociate the perception and production modalities to see how they influence each other during phonetic training. It sought to shed light on the relationship between
speech perception and production, and how this connectedness is related to the learning of second language phonemes. Each research question (RQ) that guided this study was investigated in its own experiment with its own group of participants. They are as follows:

Experiment 1

RQ1. Does perception-only training of two second language vowels improve perception of the same sounds?

RQ2. Does perception-only training of two second language vowels improve the production of the same sounds?

Experiment 2

RQ1. Does production-only training of two second language vowels improve production of the same sounds?

RQ2. Does production-only training of two second language vowels improve perception of the same sounds?

RQ3. Does access to the auditory feedback loop (for the Production-Only group) versus no access (for the Production-Only, No Sound group) influence the development of second language speech perception and production after production-only training?

Experiment 3

RQ1. How comparable and useful are bilingual and native speaker baseline groups?

The perceptual training procedure in Experiment 1 modeled the successful inhibitory training in Kondaurova and Francis (2010), while the production training in Experiment 2 combined various successful investigations that utilized computer software to provide visual, real-time feedback of pronunciation (Dalby & Kewley-Port, 1999; Ertmer, Stark, & Karlan,
1996; Flege, 1988; Franklin, 2009; Frostel et al., 2011; Lord, 2005; Massaro & Light, 2004; Osberger, 1987; Paganus et al., 2006; Ryan, Madsen, Buckley, Alex, & Bronstein, 2014; Saito, 2011; Wik & Escribano, 2009). However, and crucially different from past research, the perception training (Experiment 1) offered no opportunities for production practice, and the production training (Experiment 2) offered no opportunities for auditory input. Both perception and production were tested before and after the training in both cases. Using this two-pronged approach, I examined if improvements in one modality affected the ability of the other, providing evidence for the interdependence or independence of second language perception and production processing of two L2 vowel segments.

Some more explanation of the rationale for the design of production training in the present dissertation is warranted here, since this element and its operationalization into two distinct Production-Only trainings are the most novel contribution of the present research.

In choosing the design of a production training program in the absence of any aural input, I first reviewed literature in the deaf and hard-of-hearing population, which has an established body of work in this line of research. The most effective production training programs in this genre of literature consistently utilized audio-visual stimulus and immediate feedback. For example, Massaro and Light (2004) developed a computer-animated talking head called Baldi to train seven hard-of-hearing adolescents in perception and production. The software provided articulatory information from multiple points of view. In one display, participants could see through translucent skin to view the gestures associated with the teeth, tongue, hard palate, and velum. The researchers explain that the visual input provided complementary information, so that the students could better perceive and practice productions of the difficult sounds. The study focused on consonants and consonant clusters, and the students improved accuracy of perceiving
and producing these sounds. Rather than providing participants with visual information of the articulators, a number of researchers have used real-time spectrograms to successfully train hard-of-hearing participants’ production of segments (Ertmer & Maki, 2000; Ertmer et al., 1996; Stark, 1971, 1972). Schmidt and Beamer (1998) explain that these types of training programs provide “visual feedback to normally invisible events” (p. 389).

The majority of production training studies provide participants with an auditory stimulus to mimic or repeat. As I noted in the literature review, Herd (2011) developed a production training program and explicitly stated that the participants did not hear an auditory prompt. Instead, subjects were asked to visually compare the waveforms of their productions with those of a native speaker. Approximately thirty minutes of the first training session were used to explain how to identify and differentiate the features of [r], [ɾ], and [ð] in a waveform display. This type of task requires a certain level of expertise in order to complete the training. Although Herd was successful in creating a training program without the use of auditory stimuli, this type of complex task is not expected to be intuitive for all populations of language learners. Furthermore, in that study the participants trained in this production training task did not significantly improve their productions.

In the absence of aural input, production training programs have found success with visual images and real-time feedback. In the present study, I provided participants with both of these elements through a computer program called Vowel Shapes, an open-source computer program developed by Ryan et al. (2014). Importantly, the visual images that participants received were not arbitrary representations of their speech. Rather, the representation of their vowels was a real-time expression of their F1 and F2 values (in other words, tongue height and
backness), which means that participants were able to use this information to adjust their tongue and mouth positions to reach the target sound instantaneously.
Chapter 5: Method

This dissertation comprises three interrelated experiments. In the first experiment, I trained Spanish-speaking participants in the perception mode on the English /i/ and /ɪ/ vowels. In the second experiment, I trained Spanish-speaking participants in the production mode on the same two vowels. The third experiment compared the performances of two groups of participants in order to more closely evaluate the appropriateness and usefulness of using them as baselines for Experiment 1 and 2. The first portion of this Method Chapter describes four features that are the same across all three experiments: recruitment, participants, materials, and coding and analysis. The subsequent sections detail the unique materials and procedures utilized in each of the three experiments.

5.1. Participants

One hundred and fifteen adult native speakers of Spanish and twenty native speakers of English were recruited to join this study. The research had been approved by the Institutional Review Board of Georgetown University (ID #2015-0186), and all participants received an oral informed consent form and orally agreed to participate in the study. Each was paid ten dollars per contact hour for participation in the necessary tasks for the group and condition to which they were randomly assigned. Table 5 contains a summary of demographic information collected through the online survey for all participants in the dissertation, organized by the six groups that ultimately informed the design of the three experiments. The details are explained below.

The native speakers of Spanish (n=114) recruited for the study were between the ages of
Table 5

Participant demographic and linguistic background information

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Sex</th>
<th>Age</th>
<th>Age of English Onset</th>
<th>Years Lived in an English-speaking Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perception-Only</td>
<td>15</td>
<td>14, 1</td>
<td>31.08 (6.59)</td>
<td>10.54 (6.46)</td>
<td>1.83 (1.71)</td>
</tr>
<tr>
<td>Production-Only</td>
<td>14</td>
<td>7, 7</td>
<td>30.79 (7.99)</td>
<td>13.07 (9.01)</td>
<td>2.73 (3.15)</td>
</tr>
<tr>
<td>Production-Only, No Sound</td>
<td>15</td>
<td>12, 3</td>
<td>26.60 (4.84)</td>
<td>11.21 (5.10)</td>
<td>1.54 (1.52)</td>
</tr>
<tr>
<td>Control</td>
<td>16</td>
<td>12, 4</td>
<td>28.55 (5.15)</td>
<td>12.00 (8.41)</td>
<td>1.44 (1.46)</td>
</tr>
<tr>
<td>Bilingual Acquired</td>
<td>16</td>
<td>7, 9</td>
<td>24.50 (5.22)</td>
<td>6.33 (3.02)</td>
<td>8.47 (8.78)</td>
</tr>
<tr>
<td>NS English</td>
<td>20</td>
<td>15, 5</td>
<td>26.90 (5.16)</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Note. Means are presented with standard deviations in parentheses.

eighteen and forty-five. They were international students and professionals living in Tampa, FL and Washington, D.C. and hence highly skilled users of English who demonstrably were capable of meeting academic and professional demands requiring the advanced use of English on a daily basis. All were either currently taking or had previously taken English-medium, college-level courses or were holding professional positions in which they were expected to use English on a regular basis. Potential participants saw a recruitment flyer in person or online, or were recommended to join the study by word of mouth (see Appendix A for the recruitment flyer).

The flyer indicated that I was searching for Spanish speakers who wanted to improve their English pronunciation, and indeed 95% of the 114 NSs of Spanish who eventually participated in the study stated they had a desire to improve their English pronunciation. If potential participants asked for more information by email, I indicated that the study would be multiple thirty-minute, one-on-one sessions in which they would do listening and speaking activities that would focus on a few difficult English sounds. After contacting me by email, potential participants were directed to fill out an online background questionnaire. This form collected information about potential
participants’ age, nationality, linguistic background, years of education, length of residence in an English speaking country, daily L1 and L2 use, and self-perception of and attitudes towards their pronunciation abilities. The background questionnaire also asked participants if they had normal healthy hearing. I confirmed that a potential participant was eligible to join the study (based on age, language, proficiency, and hearing requirements), and then arranged a day and time for the participant to come to the lab (or meet at another quiet location) to complete the first session. This meeting served as a screening session to determine eligibility to participate in the study in addition to becoming a pre-test, for those who were eligible and thus continued their participation and completed the full study. Because this first session was used to determine final participant recruitment into the study and membership into the groups shown in Table 5, I describe it in detail here.

The first session was thirty minutes and consisted of a series of perception and production tests targeting the English /i/ and /ɪ/ vowels. The fact that the testing focused on these vowels was not hidden from participants, and by the end of the session they knew these vowels would be the focus of the study. As mentioned, the function of this first session was two-fold, as a screening and pre-test session. First and foremost, it was used as a method of screening participants into an appropriate group. Depending upon performance on one perception (i.e., Identification with three types of stimuli) and one production (i.e., Elicited Production) task, participants would either qualify or disqualify to continue with more sessions; both of these tasks will be described in detail in Section 5.2. The two tasks gave me information about participants’ perceptual and productive abilities of English /i/ and /ɪ/. In order to be invited to continue on and participate in one of the training regimens or as a control, the L2 subjects had to score lower than 80% accuracy on both the perception and production tasks. The Identification task score was an
average accuracy score across three types of stimuli (i.e., two sets of synthesized tokens and one set of naturally recorded tokens). To score the Elicited Production task, I listened to participants’ production of nine minimal pairs of words containing /i/ and /ɪ/. I identified which vowel I heard and checked it against the word that the participant was prompted to say. The score on this test was also an accuracy score. These screening criteria are comparable to previous perception and production training studies (e.g., Bradlow et al., 1999; Hardison, 2003; Hazan et al., 2005; Iverson et al., 2012; Lambacher et al., 2005; Wang, 2002), and they helped to distinguish if participants had established categorical knowledge of the vowels and thus could potentially benefit from training.

Sixty L2 participants were found to score below 80% on both tasks, and thus were eligible to continue with more sessions. They were randomly assigned to one of four groups with consideration of willingness and availability to complete further sessions: perception-only training (n=15), production-only training (n=14), production-only training with no sound (n=15), and control (n=16). Sixteen participants turned out to score above 80% on both perception and production screening tests, and therefore did not qualify to continue with more sessions. Because of their high scores in both modalities, which is an indication that they had acquired the two English vowels, this group’s performance at the pre-test session were retained for the study in order to constitute a baseline group. They will henceforth be called the “Bilingual Acquired” group. Their test data was retained to establish a point of comparison as bilinguals who had achieved a categorical separation of the English /i/ and /ɪ/ vowels. An additional thirty-nine NSs of Spanish did not qualify to continue with more sessions, because (a) they scored above 80% on the perception task, and below 80% on the production task (n=12); (b) they scored above 80% on
the production task, but below 80% on the perception task \((n=16)\); or (c) they qualified to participate in the study, but were not able or chose not to continue with further sessions \((n=11)\).

The group of twenty native speakers of English (the “NS English” group) were recruited so as to obtain NS baseline data on the perception and production tests. The participants were functionally monolingual English users, defined as adults who had a predominantly monolingual upbringing and who did not use a second language to an extensive level in their daily adult lives (Ortega, 2014). They were also recruited through a flyer or word of mouth. Just as the Bilingual Acquired participants, this group of NSs of English only came to the lab once to complete the entire first session. That is, just as the Bilingual Acquired group, the NS English group did not complete any training or post-test sessions.

The NSs of Spanish, from all participant groups, came from Spain, Mexico, Colombia, Costa Rica, Venezuela, Cuba, Peru, Argentina, Panama, Uruguay, Chile, Bolivia, Puerto Rico, Guatemala, and the Dominican Republic. The trained and control groups began learning English on average between the ages of ten and thirteen years old, and lived in an English speaking country on average between one and a half to three years. The Bilingual Acquired participants were all native speakers of Spanish who began learning English at an average age of six and had lived in an English speaking country at the time of testing for an average of eight and a half years. The NSs of English grew up in Florida, Pennsylvania, California, Virginia, Massachusetts, Indiana, Colorado, New York, North Carolina, and Kentucky. They had monolingual childhoods and professed to not have high proficiency in a second language nor did they extensively use an L2 in their daily lives.
5.2. Materials

5.2.1. Target vowels and auditory stimuli

The phoneme targets of this dissertation were two English vowels, /i/ and /ɪ/. I created the auditory stimuli needed for the various perception tests and the perception-only training task following the detailed descriptions in Kondaurova and Francis (2010). Speech samples were recorded with a Zoom H4N Pro Handy Recorder from a male NS of English with a standard American English dialect. In a sound attenuated booth, he was recorded saying multiple tokens of *sheep* and *ship* (shVp), and *feet* and *fit* (fVt), to be later manipulated synthetically. Additionally, he was recorded saying multiple tokens of eight /i/-/ɪ/ minimal pairs; henceforth, this set of words are called the “naturally recorded tokens”. Perception-only training task stimuli were derived from the sheep-ship tokens only, pre- and post-test stimuli requiring a discrimination response also featured the sheep-ship tokens only, and pre- and post-test stimuli requiring an identification response included the sheep-ship, feet-fit, and naturally recorded tokens. In every instance that shVp and fVt sets of stimuli were used, the tokens were synthetically altered from the original recorded samples; this process is explained in detail below. The remaining naturally recorded tokens were not altered in any way from the original recorded samples, with the exception of adding twenty-five milliseconds of silence before the token and seventy-five milliseconds of silence after the token.

Appendix B lists all of the /i/-/ɪ/ minimal pairs (i.e., shVp, fVt, and naturally recorded tokens) and accompanying depictions that were used in various tasks. This set of words represents a variety of parts of speech, but part of speech is not necessarily balanced within a minimal pair (e.g., *deep* and *dip*). All words were selected because they were high frequency and could be easily depicted in an illustration, which was necessary for the Identification test and
training task. The minimal pairs were balanced for oral frequency using data from the Corpus of Contemporary American English (COCA). This set of minimal pairs was chosen from a larger pool because each word was closest to its minimal-pair counterpart in oral frequency (/i/ to /ɪ/ word) and the pair had the highest overall frequency rankings out of the larger list, which offered the best chance for participants to be familiar with the words. An independent samples $t$-test showed that the /i/ and /ɪ/ word list oral frequencies were not significantly different from each other, $t(18)=-1.59, p=.13$.

In order to synthesize the vowels for the sheep-ship and feet-fit tokens, one sample of each word was selected for clarity from the native speaker recordings, and the onsets and codas were preserved for use throughout the study (the word final stop was fully released in all of the words). Using Praat, one vowel portion was selected for extraction from the recordings, filtered, and synthetically altered along twenty-nine formant steps (adjusting F1 and F2) to fall within the range of the /i/-/ɪ/ continuum between the vowel contrasts. Four steps equaled one just noticeable difference (JND)$^4$ apart for NSs of English (Iverson & Kuhl, 1995; Kewley-Port & Watson, 1994; Kondaurova & Francis 2010; Salminen, Tiitinen, & May, 2009). Step 1 represented a prototypical /i/ and Step 29, the prototypical /ɪ/. I determined prototypical vowel formant values by averaging /i/ and /ɪ/ productions from eleven adult male American English speakers taken from the Speech Accent Archive. For the /i/ vowel, the average F1 was 341 Hz and F2 was 2034 Hz. For the /ɪ/ vowel, the average F1 was 429 Hz and F2 was 1744 Hz.

Each of the twenty-nine spectrally altered vowel steps was next adjusted to have eleven duration steps. I measured the natural vowel lengths from the native speaker recorded samples for both vowels (mean /i/ duration=132 ms, and mean /ɪ/ duration=105 ms). Because these

$^4$ JND is used to describe the least change necessary for listeners to perceive a difference.
Figure 2. Grey boxes represent all of the synthesized vowel tokens that were created along spectral and duration continuums. The grey boxes in (A) represent all of the seventy created tokens, (B) depicts the tokens that were used for identification training, (C) depicts the tokens used for identification testing, and (D) depicts the tokens used for discrimination testing.
durations were similar to the ones found in Kondaurova and Francis (2010), I decided to maintain the duration intervals set in the previous study. Thus, the duration steps ranged from approximately seventy to two hundred milliseconds, with each step being 12.75 ms apart. Two duration steps equaled 1 JND. Steps 1, 2, 3, and 9, 10, and 11 were exaggerated tokens at the short and long end of the duration spectrum. After the vowels were adjusted for formants and duration, they were added back to the preserved onsets (i.e. [ʃ] and [f]) and codas (i.e., [p] and [t]) to form complete words. In sum, seventy synthetically altered stimuli were created and preserved for the sheep-ship continuum and seventy synthetically altered stimuli were created and preserved for the feet-fit continuum for use in this study. Figure 2 depicts the synthetic tokens used for each task.

5.2.2. Pre- and post-tests

5.2.2.1. Perception tests

The perception tasks used at pre-test (which also functioned as the prescreening session) and post-test consisted of one “Identification” task and one “Discrimination” task. Pre- and post-test versions of the Identification and Discrimination task were identical. For the Identification task, participants were seated at a computer in a sound-attenuated or quiet room. Through headphones, they were aurally presented with one word and simultaneously presented with two pictures on the computer screen. For example, the auditory stimulus sheep was played, and a picture of a sheep and a picture of a ship arranged side-by-side appeared on the screen. The presentation of the pictures remained constant throughout the task: The picture of the word containing /i/ was always on the left and the picture of the word containing /ɪ/ always appeared on the right. Participants pressed a button to select the picture that corresponded to the word they thought they heard, which was taken as an indication of the sound they perceived. Images were
used instead of written words in order to avoid an orthographic effect of the letter “i.”

This procedure with aural stimulus and picture selection was used for the synthetically adjusted stimuli (shVp and fVt) and the naturally recorded stimuli. There were 120 trials of the shVp tokens (8 spectral steps x 5 duration steps, presented 3 times each) in one block, 120 trials of the fVt tokens (8 spectral steps x 5 duration steps, presented 3 times each) in one block, and forty-eight trials of the naturally recorded tokens (8 minimal pairs, presented 3 times each) in one block. Presentation of the trials within the block was randomized; half of the correct answers corresponded to the left picture, and half corresponded to the right picture. The order of block presentations was counterbalanced across participants: the naturally recorded token block was always presented first; half of the participants completed the shVp block second and fVt block third, and half of the participants completed these blocks in the opposite order. There were 288 test trials total, which were preceded by a short practice session using the words *men* and *moon*. This was a self-paced task, and there was no time limit. The Identification task took approximately ten minutes to complete.

I decided to use the Identification task to determine if participants had categorical separation of the /i/ and /ɪ/ vowels. The three different types of stimuli were used in order to determine the effectiveness of the training on different levels of generalization. Using the shVp tokens tested the efficacy of the perception-only training with no changes to stimuli or task (same task, same vowel quality, same stimuli). Expanding the stimuli to the fVt tokens would reveal if the training generalized to a new coda and onset, while still preserving the same type of synthetic token participants heard during training (same task, same vowel quality, new stimuli).

---

5 In English, this letter read in a CVC like *sip* sounds like /ɪ/. However, it is the same letter used in Spanish to represent the vowel /i/.
Using the naturally recorded tokens as test stimuli expanded to one more level of generalization, with new onsets and codas and naturally produced words, while still preserving the task (same task, new vowel quality, new stimuli). The perception-only training would be considered most beneficial for real life conditions if participants were able to show effects on any or all levels of generalization. Crossover effects after production-only training were expected to be most easily detected in the Identification task with synthetic shVp stimuli, followed, in order, by the synthetic fVt stimuli and then the naturally recorded tokens.

The second perception test was the Discrimination task, which followed an AX design. For this task, only synthetically adjusted stimuli from the shVp set were used. Participants sat at a computer, and through headphones, they heard one word. Simultaneously, two boxes appeared on the screen with the words same and different. After a five hundred millisecond inter-stimulus delay, participants heard the second word and pressed the corresponding button according to their perceptions. Total, there were fourteen same pair trials (AA) and twelve different pair trials (AB), which were also presented in the opposite order (BA). The same trials (AA) were the same exact synthesized token recording, and the different pairs were one JND apart in the spectral continuum (i.e., four steps apart) or one JND apart in duration (i.e., two steps apart). I also created ten different pairs (CD) using tokens that were two JNDS apart in the spectral continuum (i.e., eight steps apart) or in duration (i.e., four steps apart) to use as filler trials that would offer participants some opportunity to press “different” with certainty. Each token of interest (i.e., AA, AB, BA) was played twice throughout the experiment, and trials were evenly divided into four blocks of twenty-four. Before the test began, there was a short practice session using the words men and moon. This was a self-paced task and there was no time limit; participants completed the task in approximately seven minutes. The Discrimination task was important to include in the
battery of tests because it was able to reveal if participants had sensitivity at the categorical boundary between the two vowels, even if full category separation had not yet occurred.

5.2.2.2. Production tests

All participants completed three production tests at pre- and post-test stages, modeled after the versions used in Munro and Derwing (2008) and Saito (2011). The production recordings were made with a Zoom H4N Pro Handy Recorder, and a laptop was used to deliver all tasks. All attempts were made to use a sound attenuated booth or quiet space to achieve high-quality recordings.

The first production test elicited spontaneous speech in the “Picture Description” task. Participants were instructed to describe a picture for approximately three minutes to someone who had never seen it before. The first ten /i/ words and the first ten /ɪ/ words that participants produced were later extracted for analysis. A research assistant drew two different but similar pictures that were used for this task, and the two versions were counterbalanced across participants in pre- and post-test stages; see Appendix C for the pictures. The illustrations were piloted with a small group of people to ensure that they were likely to elicit words containing /i/ and /ɪ/. However, as a precaution, each picture was embedded with fifteen items whose referent words contained /i/ and fifteen items whose referent word contained /ɪ/ (e.g., bee, piggy bank, leaves, fish); see Appendix D for a list of these words. The embedded /i/ and /ɪ/ words were balanced for oral frequency based on data obtained from COCA. Independent samples t-tests revealed that there was no difference in frequency between versions (t(58)= -.38, p=.71) nor between vowels (t(58)= -.30, p=.76).

Participants next completed a second production test, the “Passage Reading” task. They read aloud a short passage of approximately two hundred words, excerpted from Jane Austen’s
Pride and Prejudice; participants were asked to read the passage aloud two times in order to ensure the best chance for acceptable recordings of target words. Productions of the target words from the first reading were given preference for extraction because the second reading tended to be spoken faster and many vowels were reduced and shortened; productions from the second reading were only used in the case of circumstantial or technical mishap. I selected two passages that were counter-balanced across participants for the pre- and post-tests (see Appendix E), and I identified ten /i/ words and ten /ɪ/ words from each of the passages that served as test items. The /i/ and /ɪ/ word lists were balanced for oral frequency based on data obtained from COCA. Independent samples t-tests revealed that there was no difference in frequency between versions (t(38) = .22, p = .83) nor between vowels (t(38) = -.27, p = .79).

In the third and final production test subjects were recorded saying a list of eighteen single C(C)V words in the “Elicited Production” task. The list included the two trained words (i.e., sheep and ship) and eight words from the naturally recorded set; each word was repeated twice in a random order. In order to familiarize participants with the illustrations that represented these words, they first completed a short training task prior to the Elicited Production task proper. In the first portion of the training, participants clicked through images and read each picture’s associated word silently. In the second part, participants saw a picture and were prompted to type the associated word. Any missed item was repeated until the participant reached 100% accuracy on this task. It was determined that using the orthographic components in this familiarization task was unavoidable, as this task ensured that participants could associate the picture with the lexical item which would be used in the subsequent task. Crucially, during the test in which participants’ productions would be recorded, the orthographic form of the word was not presented with the illustration. Immediately after the familiarization task, participants began the
Elicited Production task. The screen displayed a picture along with the first consonant(s) of the word followed by a blank line, and participants said the word associated with the picture. Before the test trials began, there was a practice session with the words men and moon. Words from the second repetition were extracted for later analysis because it was presumed by the second half of the task participants were comfortable with the procedure and stumbled less often, which would produce a clearer token than the first round of productions.

Completion of all three production tasks took approximately fifteen minutes total. By capturing production of the target vowels in three different elicitation methods, speaker attention ranged from more to less careful (i.e., Elicited Production, Passage Reading, and Picture Description, in that order) and it could be seen if participants generalized improvements from the easiest condition of careful speech to the hardest condition of spontaneous speech. In the same way that the perception tests were designed to test generalization after training, the three production tests were designed to capture the effectiveness of training at different levels of generalization. The Elicited Production task was the most similar to the production training task. This test did not change the elicitation method nor the stimulus onset and coda, shVp (same task, same stimulus). Next, this same task was used to elicit words with new onsets and codas, different than the training stimulus; this is the first level of generalization (same task, different stimuli). The Passage Reading task tapped a second level of generalization, by using a less controlled elicitation method and new target words (new task, new stimuli). Finally, the Picture Description task was a spontaneous speech context, which represented the least careful speech elicitation method and elicited new words (new task, new stimuli). The Elicited Production task was the most likely to show effects after the production-only training, however, the production training would be most ecologically valid if it could show effects on any or all levels of
generalizations. Crossover effects after perception-only training were expected to be most easily detected in the Elicited Production task, followed, in order, by the Passage Reading task and then the Picture Description task.

5.2.3. Native speaker identifications of productions

Participants’ productions of the /i/ and /ɪ/ words were judged by three NSs of English. The researcher was the first rater (aged 33), and two additional raters were in their mid-sixties. All had normal, healthy levels of hearing, and all had extensive experience in phonetics and SLA. The three raters completed a category judgment task of all participant recordings from the three production tasks. Productions from all three tasks were spliced so that only the onset and vowel portion remained. Isolating the onset and vowel sounds was necessary because not all produced tokens were minimal pairs, and knowledge of the participants’ intended productions could undesirably influence the rater to select one vowel over the other. During the Identification task, the native speaker listeners heard one token of an onset and vowel, and they selected one of three options: /i/, /ɪ/, or OTHER. All listeners rated all tokens (approximately 9,000 tokens) twice; in other words, each participant token was rated six times. This task resulted in a percentage of accurately identified tokens for each participant on each task. In order to avoid confusion or rater fatigue, distractor vowel trials were added to the task and raters limited time-on-task to a reasonable amount per day.

This method of production rating was modeled after Munro and Derwing (2008), who stress that asking raters to provide a categorical identification upon hearing a recorded token gives an indication of a non-native speaker’s intelligibility, not a rating of his or her accent. They explain, “If a speaker’s production of the vowel in ‘bit’ sounded more like /i/ than any other English vowel then the production would be deemed intelligible, even if it was not a perfect
exemplar of /i/” (p. 485). Rating participant productions in this way reorients the field of SLA toward intelligibility and away from the “native speaker as ideal” approach. Even though this type of rating system focuses on intelligibility, scores still reflected the participants’ categorical knowledge and their ability to produce an acceptable (to a native speaker listener) token of each vowel.

The perception and production training materials differed across experiments. Therefore, each will be presented later in this chapter, under the methods specific to each experiment. I turn now to the rest of methodological explanations that apply across experiments, comprising other data collection instruments administered to all participants, and coding and analysis procedures.

5.2.4. Other data collection instruments

5.2.4.1. Background questionnaire

Participants completed an online background questionnaire through Google Forms before the pre-test session. First, basic demographic information was collected such as gender, age, country of birth, years of education, and hearing health. Second, additional questions elicited information about first and second language use, such as age of English learning onset, length of residence in an English speaking country, and daily L1 and L2 use. Finally, participants were asked about their self-perception of and attitudes towards their pronunciation abilities. The questionnaire took approximately ten minutes to complete.

5.2.4.2. Informal interviews

During and after each session across experiments, I asked the participants to share their thoughts of what was happening during the experimental sessions. I asked questions to discover what they thought was difficult or what they had noticed during the tasks and tests. Additionally, I attempted to elicit information about their behaviors outside of the sessions, with particular
focus on moments when they talked about the experiment with family, friends, or coworkers, or when they noticed the target phonemes in their regular daily lives. I also asked about their attitudes towards their pronunciation abilities in or outside of the experiment and any changes they may have perceived in their listening or speaking over the duration of the study. The information gathered during these informal interviews was used to augment the quantitative results gathered at pre- and post-tests.

5.3. Coding and analysis of test performances

Before all statistical tests were run, alpha was set at $p<.05$. The identification task was scored by accuracy according to the standards set by the NS English group. Figure 3 shows how the majority of NSs of English categorized each synthetic token of /i/ and /ɪ/ in the shVp and fVt contexts. A synthesized token was determined to be an /i/ or an /ɪ/ if more than fifty percent of the NSs identified it as such. The treatment, control, and Bilingual Acquired groups’ accuracy on the Identification tasks with synthesized stimuli were compared against the identifications of the NS English group, and a percent accuracy score was recorded. Identification scores of the naturally recorded stimuli were determined by a simple accuracy measure. A score of 50% would be indicative of random button pushing, which is a sign of no knowledge of the difference between the /i/ vowel and the /ɪ/ vowel; a score of 100% would indicate perfect ability to distinguish the two sounds and associate them with the correct lexical item; and a score of 0% would mean perfect ability to distinguish and assign the two vowels to a specific picture, just a transposed lexical assignation, in other words always assigning the [i] to ship and always assigning [ɪ] to sheep. In order to test for change from pre- to post-test, a repeated measures ANOVA was run on the experimental and control group data. A
one-way ANOVA was used to determine if the experimental and control groups were significantly different from Bilingual Acquired and NS English groups at pre-test and again at post-test.

The purpose of the Discrimination task was to determine if participants had sensitivity at a category boundary. In order to determine if participants had this sensitivity, I converted the responses of hits and false alarms to a sensitivity measure ($d'$) following signal detection theory (Macmillan & Creelman, 1991). It was expected that $d'$ scores for the bilingual and NS English groups would reveal a sensitivity spike around spectral Steps 13 and 17, and there would be no spike on the duration continuum. These data were visually inspected on a graph that depicted the
d’ score along the spectral continuum and another graph that depicted the d’ score along the duration continuum.

In order to analyze participant productions from the three production tasks, /i/ and /u/ vowels (along with their onsets) were extracted from the longer recordings. Two criteria were taken into consideration when selecting which tokens to preserve for analysis: vowels could not be followed by a nasal or liquid coda, and the vowel could not appear as part of a reduced syllable. These criteria were important because participant productions were acoustically analyzed and averaged, and these phonetic contexts greatly affect formant values.

The extracted onset and vowel tokens were analyzed in two ways. First, NS judges identified all produced vowels as described in Section 6.2.3, and this produced a simple accuracy rating. A repeated Measures ANOVA was run with the experimental and control group data to see if there was any change from pre- to post-test. A one-way ANOVA was used to determine if the experimental and control groups were different from the Bilingual Acquired and NS English groups at pre-test and again at post-test.

The productions were also acoustically analyzed. While extracting onsets and vowels from participants’ speech recordings, I took care to mark the end of a vowel before the major portion of the coarticulation of the coda began. Thereafter, a Praat script extracted formant values for all vowels at the 80% point of the full vowel length (a likely steady state). Next, a vowel overlap calculator created by Daniel Ezra Johnson (Johnson, 2016) was used to identify the percentage of overlap between the two vowel categories for each participant. There was one overlap score per participant at pre-test and another at post-test. Thus, all three production tasks were combined for one pre-test score, and again all three were combined for a one post-test score. It was necessary to combine the data in this way because the overlap measure is not reliable with
small numbers of tokens. By combining all three tasks at pre- and again at post-test, each participant had approximately thirty tokens of /i/ and thirty tokens of /u/ that could be used for overlap analysis. The overlap percent for the experimental and control groups at pre- and post-tests were analyzed with a repeated measures ANOVA.

The second and final type of acoustic analysis I conducted was a visual inspection of group mean /i/ and /u/ pre- and post-test productions separated by task and plotted on a vowel space. In order to plot the group mean vowel productions, I first had to Bark normalize all formant values. Each human’s unique vocal anatomy creates unique formant values spaced out on unique vowel planes. Thus formant values cannot be compared across individuals, and a normalization method must be used. Normalizing the formant values to Bark z-scores helps to mitigate individual differences and vowels can then be compared across participants. Once formants are transformed to z-scores, the Z3-Z1 score becomes the equivalent to F1. On a vowel plot, this is the y-axis, which determines vowel height. The Z3-Z2 score is the equivalent to F2. On a vowel plot, this is the x-axis, which is the vowel backness (or fronting). In a normal vowel plot with F-values, the origin is in the upper right corner. However, with Bark values, the origin is in the lower left corner. This means that the higher the Z3-Z1 score is, the higher the vowel is, and the lower the Z3-Z2, the more fronted the vowel is. Once mean vowels were plotted on a vowel space, the figures were visually inspected to determine if the mean /i/ or mean /u/ moved closer to target-like productions by post-test.

5.4. Experiment 1

The objective of Experiment 1 was to determine if the production system is accessed and affected by the perception modality. The perception-only training utilized in this study was modeled after the procedure in Kondaurova and Francis (2010). These authors targeted NSs of
Spanish who showed a reliance on duration rather than formant cues when distinguishing the English /i/-/ɪ/ contrast. Because the perceptual training of their “Inhibition Group” led to native-like identification after only two hours of training, the procedure of the Inhibition Group was used as a model in the present study. I extended the previous study by adding a production pre- and post-test, and I also expanded the sample population by offering the training to all native speakers of Spanish who showed difficulty categorizing /i/ and /ɪ/ instead of just targeting those who categorized the two vowels based on duration cues. Thus, in the present experiment, fifteen NSs of Spanish underwent perception-only training and sixteen acted as controls.

5.4.1. Perception-only training

The training task in Experiment 1 was similar to the Identification task employed in the pre- and post-tests, with a few essential differences. Participants who underwent this training (the “Perception-Only” group) heard a presentation of a word on the sheep-ship continuum, and they chose the picture that corresponded to what they thought they heard. Synthetic shVp tokens from the ends of the spectral continuum were utilized for the training: Vowels from Steps 1 and 5 were used for the sheep tokens, and vowels from Steps 25 and 29 were used for the ship tokens. All manipulations along the duration continuum were used for both vowels (Steps 1-11). (See Figure 2B for a visual representation of the synthetic tokens used in the perception-only training task.) Unlike the Identification task used in the testing conditions, participants received feedback in the Identification task used in the training condition. There was a three second time limit for a participant to respond, and after an answer had been selected or time had elapsed, feedback appeared on the screen: a green checkmark for a correct answer, or a red X for an incorrect answer or no selection. When the participant pushed the button to continue, the audio stimulus was played again with only the correct picture displayed on the screen. The next screen presented
how many correct trials the participant had completed, and then a burst of white noise (1200 ms) sounded to eliminate the effects of echoic memory from the previous trial. The next stimulus was then presented. There were 132 trials, divided into four blocks; twenty-two of the synthesized /i/ tokens and twenty-two of the synthesized /ɪ/ tokens were presented three times each. Thus participants heard 264 iterations of the /i/ and /ɪ/ vowels in one training session. One session lasted approximately thirty minutes, and participants completed all four sessions within a two to three week time period.

5.4.2. Procedure

Table 6 summarizes the procedure for the experimental group only. Before the first session, potential participants filled out an online background questionnaire, which served two purposes. The first was to ensure that the potential subject qualified for the experiment, and the second was to collect demographic and linguistic information. After it was determined that a subject was eligible, he or she came to the lab (or met the researcher in another quiet location) for the pre-test/prescreen session, which lasted approximately thirty minutes. Participants completed the following tasks in a fixed order: three Production Tasks (i.e., Picture Description, Passage Reading, and Elicited Production), one Identification Task with three types of stimuli (i.e., synthetic shVp, synthetic fVt, and naturally recorded tokens), and one Discrimination Task with one set of stimuli (i.e., synthetic shVp tokens). It was important that the production tasks preceded the perception tasks so that the auditory stimuli would not influence participants’ pronunciation of the words. The experimental group returned to the lab (or another quiet location) four times to complete thirty-minute identification training sessions, for a total of two hours of training. After four training sessions, participants returned to the lab to complete the
Table 6

Description of testing and training days for the Perception-Only experimental group

<table>
<thead>
<tr>
<th></th>
<th>Pre-test</th>
<th>Training 1</th>
<th>Training 2</th>
<th>Training 3</th>
<th>Training 4</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 minutes</td>
<td>30 minutes</td>
<td>30 minutes</td>
<td>30 minutes</td>
<td>30 minutes</td>
<td>30 minutes</td>
<td>30 minutes</td>
</tr>
<tr>
<td>Production Tests</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Picture Description</td>
<td>Identification</td>
<td>Identification</td>
<td>Identification</td>
<td>Identification</td>
<td>Identification</td>
<td></td>
</tr>
<tr>
<td>Passage Reading</td>
<td>Identification</td>
<td>Identification</td>
<td>Identification</td>
<td>Identification</td>
<td>Passage Reading</td>
<td></td>
</tr>
<tr>
<td>Elicited Production</td>
<td>Task with Feedback</td>
<td>Task with Feedback</td>
<td>Task with Feedback</td>
<td>Task with Feedback</td>
<td>Elicited Production</td>
<td></td>
</tr>
<tr>
<td>Identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrimination</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Post-test session. This included all of the tasks given at the pre-test session, the only changes being the version of the Picture Description and the Passage Reading tasks; versions of the tasks were counterbalanced across participants. Experimental group participants completed pre-test, training, and post-test over a two to three week time period, and I strongly encouraged participants to complete the last day of training and the post-test session on consecutive days, although this was not possible for all participants. The Control group completed the same pre-test session and returned to the lab approximately two to three weeks later to complete the post-test session. The Bilingual Acquired and NS English baseline groups only completed the pre-test session and did not return for more sessions.

5.5. Experiment 2

The objective of Experiment 1 was to determine if the production system is accessed and affected by the perception modality. Experiment 2 was designed to ask the opposite question: Can the perceptual system be accessed and affected by the production modality? In order to truly isolate the production mode, participants receiving production-only training could not be exposed to any aural tokens during the course of their training. In order to achieve this, two
experimental groups were created in Experiment 2. The first was a production-only training group (“Production-Only”), who completed production training in individual sessions, using an open-source sound visualizer program and never heard other-produced aural examples of the target phonemes. However, this group was able to hear their own voices. This counted as self-produced auditory feedback, creating a possible perception-training element. The second experimental group therefore also underwent the same visual training, but in addition they were not able to hear their own voices (“Production-Only, No Sound”). Because this study highlights production training that occurred in the absence of all auditory input, it was desirable to create a training group that would truly hear no aural presentations of the target stimuli, even from their own voices. In the strictest sense, this group did not hear any iterations of the target sounds during training from which it could be said that perception was incidentally being activated and trained. Fourteen participants were in the Production-Only group, and fifteen participants were in the Production-Only, No Sound group. Data from the Control, Bilingual Acquired, and NS English groups were again used as points of comparison.

5.5.1. Production-only training task

The program that was used for the present experiment’s production-only training for both experimental groups is called Vowel Shapes. In 2014, a group of graduate students at the University of Rochester developed an open-source computer program that aided vocal students in improving the quality of their sung vowels (Ryan et al., 2014). Although the program was originally created for students of music, it is easily adapted to second language learners as well. The program visually depicts normalized F1 and F2 values on multiple interfaces. In the graph interface, a teacher or researcher can record and save a vowel as a target. The program will represent this vowel as a fixed blue dot on a plane that simulates the vowel space chart familiar
Figure 4. The Vowel Shapes visualizer. The blue dot represents the target vowel and the red dot represents the participant's production.

to linguists, where F1 is plotted on the y-axis and F2 is plotted on the x-axis. Next, a learner will say a vowel held for an extended amount of time, in an attempt to reproduce the target vowel. The learner’s vowel is represented by a red dot, which moves in real-time on the screen. The position of the dot changes as the learner alters his or her vowel production. Once the learner reaches the fixed target vowel, within an acceptable range, the red dot turns green. See Figure 4 for a depiction of the Vowel Shapes graph interface.

Pilot testing revealed that the vowel space chart was more intuitive than the other interface options for participants to understand and to make changes to their vowel productions. Additionally, it was discovered in pilot testing that target vowels (i.e., the position of the blue dot) had to be individualized for each participant. Although the Vowel Shapes program normalizes all vocal inputs, each person has a unique composition of vowels in the vowel space. Thus a target-like /i/ for one participant is not in the same spot as a target-like /i/ for another. Therefore, a portion of time on the first day of training was used to find where each participant’s target /i/ and target /ɪ/ needed to be located for subsequent training. The final detail to emerge from pilot testing was the percent of acceptability that provided participants with the largest
accommodation zone while still producing an acceptable target sound. One detail that went into consideration for choosing this number was the close proximity of the /i/ and /ɪ/ vowels. In specific terms, a twenty percent zone above the /i/ target was still an acceptable /i/, but twenty percent below the /i/ target could become an /ɪ/ sound. Real vowel categories are not perfect circles around one ideal origin. Rather, vowel category shapes are irregular and nebulous. Ultimately, an eighteen percent acceptability zone seemed to provide the best accommodation zone under these constraints.

The procedure for the production-only training task for both experimental groups was similar to the Elicited Production task employed in the pre- and post-tests, in that participants saw a picture and said the corresponding word. However, there were other substantial changes. In the training, participants sat in front of two laptop computers: The left laptop screen displayed instructions for the Elicited Production task and the image prompts for each trial, while the right laptop displayed the Vowel Shapes program. On the left laptop, the following phrase was displayed at the top of the screen: Say this word now, and extend the vowel for X seconds. Underneath the phrase, a picture of a sheep or a ship appeared simultaneously with the phrase. To the right of the picture, a clock counted down the number of seconds participants were to hold the vowel sound. Participants were instructed to extend the vowel for two, four, six, eight, or ten seconds, and at the end of the time, the countdown clock said STOP. Participants spoke into a microphone, saying the target word with an extended vowel sound, and watched the right screen with the Vowel Shapes program (see Figure 5 for a mock-up of the two screens). Within the time frame for one trial or across multiple trials, they could adjust the sound of the vowel in order to move the red dot to reach the blue target. After one trial at a specific duration, participants were instructed to try one more time at that same vowel length.
Next, a screen appeared that directed them to stick out their tongues and relax their mouths for five seconds; then the next trial began.

The first day of training began with an introduction to the Vowel Shapes program and an explanation of how it worked (see Appendix F for the instructions participants received). Next, participants practiced with the program, extending the vowel sound and manipulating it to control the movements of the red dot. This practice offered the researcher time to determine the best location for each participant’s ideal target vowel (i.e., the blue dot). Finally, the participants practiced the full production training task for a shortened period of time. On this first day of training, participants practiced all five duration variants with an immediate repetition of each, for both the /i/ and /ɪ/ vowels. This totaled forty trials, which were evenly divided between four blocks. The presentation of trials and the vowel target of each block were randomly selected before the beginning of the experiment. On the second day of training, participants studied a handout (see Appendix G) that showed images of the mouth shape and a diagram of the internal and external articulator positions for each vowel. On this second day, the individualized target vowels had been set, and full training began. See Table 7 for a description of each training day.
One day of training lasted approximately thirty minutes. Before or after participants completed the standard trials, they were given time to practice freely with the Vowel Shapes program. Additionally, I encouraged participants to share their thoughts, insights, and challenges as they completed the training.

As mentioned, there were two experimental groups in Experiment 2. The Production-Only group completed the production-only training with the Vowel Shapes visualizer in individual sessions in a laboratory or another quiet space. This experimental group received absolutely no other-produced aural input during the training sessions, although they were able to hear their own voices. The second experimental group, the Production-Only, No Sound group, never received other-produced aural input during the training, and they were not able to hear self-produced auditory feedback as well. This second experimental group completed the same tasks as the Production-Only group, with one exception. Participants in the Production-Only, No Sound group wore Bose QuietComfort 25 Acoustic Noise-Cancelling headphones to block the acoustic information from their own productions that travel through the air, also called air-conducted sound. Even with the air-conducted sound canceled, a person can still hear him- or herself through bone-conduction (Stenfelt & Goode, 2005). Natke and Kalveram (2001) have canceled out this type of sound for participants by playing 900 Hz low pass filtered white noise at 70 dB through headphones. This experiment utilized similar techniques, and played participants a combination of low pass filtered white noise (i.e., pink or brown noise) and rain sounds that were suitable for canceling out their bone-conducted speech. Before training began, the Production-Only, No Sound participants talked with the headphones on, and together with the researcher decided which color of noise and what volume was sufficient to block the sound of their voices. The United States Department of Health and Human Services, in conjunction
with the Center for Disease Control and Prevention and the National Institute for Occupational
Safety and Health, developed guidelines for safe hearing standards in the workplace. The
standards suggest volume levels and the amount of time that workers can safely be exposed to a
sound. Using these guidelines, the dB level of the noise never exceeded 97 dB in a thirty-minute,
continuous time period. As another precaution, the training task had built in breaks to give
participants a chance to rest in between blocks of trials if listening to the noise for long periods
of time became frustrating or tiresome.

I made every effort to maintain comparability between the perception-only training in
Experiment 1 and the production-only training in Experiment 2. The very nature of the
modalities required that some aspects of the training would be different. For example, the
perception training of Experiment 1 used tokens of /i/ and /ɪ/ that were synthesized to eleven
duration steps, each step separated by 12.75 ms. It was not possible for the participants in the
production training study to mimic duration steps at such small increments. Therefore, the
variability of vowel length in the production training was set at much larger intervals: two, four,
six, eight, and ten seconds long. This, in turn, directly affected the density of stimuli in each of
the trainings. In the perception training, participants were able to listen to an aural stimulus and
push a button to identify the sound within five seconds. On the other hand, one trial of
production training could last three times as long.

Just as there were some features that needed to be different between the two training
programs, there were many features that I was able to hold constant across the two experiments.
Table 8 lists the substantive features of the two training programs and shows how they were
similar or different. Although exact comparability was not achieved, the main goal of both
Table 7

*Description of trials and blocks for each day of production training*

<table>
<thead>
<tr>
<th>Training Day</th>
<th>Total Number of Trials</th>
<th>Order of Blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
<td>1 - i - 1 - i</td>
</tr>
<tr>
<td>2</td>
<td>56</td>
<td>i - 1 - i - 1</td>
</tr>
<tr>
<td>3</td>
<td>56</td>
<td>1 - i - 1 - i</td>
</tr>
<tr>
<td>4</td>
<td>48</td>
<td>1 - i - 1 - i</td>
</tr>
</tbody>
</table>

Table 8

*Comparability of features in the perception-only training of Experiment 1 and the production-only training of Experiment 2*

<table>
<thead>
<tr>
<th>Substantive Feature</th>
<th>Perception Training</th>
<th>Production Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrier Onset and Coda</td>
<td>shVp</td>
<td>shVp</td>
</tr>
<tr>
<td>Stimuli</td>
<td>/i/ and /ɪ/ at 11 duration steps per vowel</td>
<td>/i/ and /ɪ/ at 5 duration steps per vowel</td>
</tr>
<tr>
<td>Variability</td>
<td>Duration</td>
<td>Duration</td>
</tr>
<tr>
<td>Feedback</td>
<td>Checkmark or X</td>
<td>After reaching blue target, circle turns green</td>
</tr>
<tr>
<td>Feedback 2</td>
<td>Replay of the stimulus with correct picture</td>
<td>Immediate repetition of word at same duration</td>
</tr>
<tr>
<td>Number of Sessions</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Length of Each Session</td>
<td>30 minutes</td>
<td>30 minutes</td>
</tr>
<tr>
<td>Number of Stimulus Presentations</td>
<td>264 per session</td>
<td>40-56 per session</td>
</tr>
<tr>
<td>Interstimulus Break</td>
<td>Burst of white noise</td>
<td>Stick out tongue</td>
</tr>
</tbody>
</table>

experiments was to create a highly effective training program in one modality so that potential crossover effects could be observed in the second modality.

5.5.2. Procedure

The structure for this experiment was exactly the same as Experiment 1. Both production training groups came into the lab (or another quiet location) for pre-test, four days of training, and the post-test in a two-to-three week time period. Every effort was made to hold the last day
of training and the final post-test session on consecutive days, although this was not achieved with all participants.

5.6. Experiment 3

The objective of Experiment 3 was to directly and closely compare the performances of the NS and Bilingual Acquired groups, with the ultimate goal of determining if highly proficient bilinguals are as good as NSs at distinguishing a difficult vowel contrast, and if so, does that mean bilinguals are an appropriate comparison group moving forward in L2 phonetic training research. To that end, the sixteen Bilingual Acquired participants and the twenty NS English participants in the two baseline groups were compared in this experiment.

5.6.1. Procedures

Procedures for this experiment solely consisted of the pre-test/prescreen session. Participants completed three Production tasks (i.e., Picture Description, Passage Reading, and Elicited Production), one Identification task, and one Discrimination task, in that order. The session lasted approximately thirty minutes. Participants did not return to the lab for any further testing.

5.6.2. Coding and analysis

In order to compare the NS and Bilingual Acquired groups, the task data were analyzed in the same way as the other experiments. The Identification task was scored with a simple accuracy measure, and the Discrimination task was converted to a d’ sensitivity metric and visually inspected for a spike across the spectral and duration continua. The productions from the three tasks were analyzed in two ways: first they were identified by the NS judges, and second they were analyzed acoustically to determine the overlap score of the two vowels and to plot the group mean productions on vowel spaces divided by task.
Beyond the same analyses that were used in the two prior experiments, Bilingual Acquired and the NS English group data were depicted in figures to elucidate which individual Bilingual Acquired participants fell in or out of the NS English range of scores on the various perception and production tasks.
Chapter 6: Results

This Results Chapter is divided into three sections, reflecting the three-experiment structure of this dissertation. First, I will report the findings from Experiment 1, the objective of which was to determine if the production system is accessed and affected by the perception modality. Next, I will present the results from Experiment 2, which was designed to address the relationship between perception and production in the opposite direction of Experiment 1: Can the perceptual system be accessed and affected by the production modality? Finally, I will report the findings from Experiment 3, which investigated the similarities and differences between the functionally monolingual English group and the bilinguals who achieved over 80% on both the perception and production tasks at the prescreening stage. This last experiment was designed with the intention of exploring the usefulness and appropriateness of two baselines.

6.1. Experiment 1

The primary objective of Experiment 1 was to determine if the production system can be accessed through the perception modality. To that end, two research questions were posed: (1) Does perception-only training of two L2 vowels improve perception of the same sounds? and (2) Does perception-only training of two L2 vowels improve production of the same sounds? I will present the findings in turn.

6.1.1. Research Question 1: Does perception-only training of two second language vowels improve perception of the same sounds?

Fifteen participants underwent perception-only training in the present experiment. They completed two perception tasks in pre- and post-test sessions: an Identification task and a Discrimination task.
Table 9 presents the scores for each group for the Identification task at pre- and post-test. Results show that participants who underwent perception-only training improved on this task with three types of stimuli: Identification of tokens similar to the training stimuli (i.e., synthetic shVp) improved by 22%, identification of the synthetic fVt tokens improved by 17%, and identification of the naturally recorded tokens improved by 11%. In comparison, the Control group improved by 8% on the shVp tokens, 4% on the fVt tokens, and decreased by 7% on the naturally recorded tokens. See the bar graphs in Figure 6 for a visual depiction of these scores; blue and grey dotted lines represent the two baseline groups. I ran a 2 x 2 x 3 repeated measures ANOVA (Group by Time by Stimuli) with the Perception-Only and Control group data from pre- and post-tests. The assumption of homogeneity was not violated, but the assumption of normality was violated for the naturally recorded stimuli at post-test. Thus, I log-transformed the naturally recorded stimuli data before running the ANOVA. Mauchely’s test of sphericity was significant for Stimuli (\( \eta_p^2 = .03 \)) and not significant for Time x Stimuli (\( \eta_p^2 = .86 \)), thus the assumption of sphericity was violated for Stimuli, and the Greenhouse-Geisser correction was used for tests with this variable. Results revealed that the Time by Group interaction was significant (\( F(1, 29) = 5.68, p = .02, \eta_p^2 = .16, \text{power} = .63 \)), which means that the groups performed differently over time. A series of paired samples \( t \)-tests show that from pre- to post-test, the Perception-Only group significantly improved on the shVp stimuli (\( t(14) = -3.45, p = .004 \)) and the fVt stimuli (\( t(14) = -2.72, p = .02 \)), but not on the naturally recorded stimuli (\( t(14) = -2.31, p = .07 \)). A paired samples \( t \)-test for the Control group revealed that they significantly improved on the shVp stimuli (\( t(15) = -2.41, p = .03 \)), but they did not improve on the other two stimuli set (fVt, \( t(15) = -1.77, p = .10 \); naturally recorded, \( t(15) = 2.01, p = .06 \)). Although the \( t \)-test suggests that the Control
Table 9

Mean percent accuracy (SD) for three types of stimuli in the Identification task for Perception-Only, Control, and baseline groups

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>shVp Pre (SD)</th>
<th>Post (SD)</th>
<th>fVt Pre (SD)</th>
<th>Post (SD)</th>
<th>Natural Tokens Pre (SD)</th>
<th>Post (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perception-Only</td>
<td>15</td>
<td>50.17 (11.82)</td>
<td>71.89 (16.41)</td>
<td>48.90 (16.53)</td>
<td>66.00 (19.99)</td>
<td>65.84 (16.98)</td>
<td>76.95 (18.95)</td>
</tr>
<tr>
<td>Control</td>
<td>16</td>
<td>44.69 (13.87)</td>
<td>52.67 (18.91)</td>
<td>50.47 (21.22)</td>
<td>53.75 (23.10)</td>
<td>69.01 (19.92)</td>
<td>62.07 (22.46)</td>
</tr>
<tr>
<td>NS English</td>
<td>20</td>
<td>85.88 (5.98)</td>
<td>-- a</td>
<td>88.63 (5.32)</td>
<td>--</td>
<td>98.54 (1.92)</td>
<td>--</td>
</tr>
<tr>
<td>Bilingual, Acquired</td>
<td>16</td>
<td>81.72 (6.17)</td>
<td>-- a</td>
<td>83.59 (8.74)</td>
<td>--</td>
<td>97.27 (2.25)</td>
<td>--</td>
</tr>
</tbody>
</table>

"Baseline groups provided pre-test data only; they were not tested again in a post-test session.

Table 10

Mean percent accuracy (SD) for the Perception-Only training group for the two trained vowels on each day of training

<table>
<thead>
<tr>
<th>Session</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
</tr>
<tr>
<td>Session 1</td>
<td>81.92</td>
<td>76.36</td>
<td>88.28</td>
<td>83.89</td>
</tr>
<tr>
<td>Session 2</td>
<td>14.07</td>
<td>14.14</td>
<td>14.08</td>
<td>11.77</td>
</tr>
</tbody>
</table>

76
Figure 6. Identification pre- and post-test scores for the experimental Perception-Only (Per) and Control (C) groups on all three types of stimuli. NS English (E) and Bilingual Acquired (B) group data are presented with dotted grey and blue lines. A red star indicates a statistically significant difference from pre- to post-test.
group improved on the shVp stimuli, the pre-to-post-test scores only changed from 45% to 53%, which still hovers around 50%, a score that is indicative of random guessing. By contrast, the training group changed from 50% to 72%. The essential information from these analyses is that the Perception-Only group statistically significantly improved on the Identification perception task with the shVp stimuli, which is most similar to the training, as well as with the fVt stimuli, which is evidence that learning generalized to a new phonetic context.

An inspection of the Perception-Only training group’s performance during training confirms the results of the pre-to-post-test improvement on the Identification task. In order to better interpret the training session performance reported here, it must be appreciated that participants were asked to listen to a sound and push a button to select which sound they thought they heard. If the participants could not distinguish the difficult /ɪ/ vowel, there would be evidence of random button pushing, and the resulting score would be close to 50%. On the other hand, a 0% score on the perception training task would mean perfect identification of the vowel sounds, except that the participant had transposed the /i/ and /ɪ/ sounds to the wrong lexical item (i.e., always marking [i] as “ship”, and always marking [ɪ] as “sheep”). Recall also that the training solely included stimuli from the ends of the spectral continuum, Steps 1 and 5 for the /i/ vowel and Steps 25 and 29 for /ɪ/. Thus, the auditory stimuli were good exemplars of the two vowels and were not intended to be ambiguous in terms of spectral cues. Rather, participants heard the full range of duration stimuli (Steps 1 through 11) in order to encourage them to rely less on this cue.

Table 10 shows the mean percent accuracy for each vowel over the course of the four training sessions. Mean scores show that participants consistently improved in their performance from one session to the next for both vowels. The percentage of correctly identified /i/ tokens
began at 82% at Session 1 and improved to 93% by Session 4. Similarly, correctly identified /i/ tokens improved from 76% at the first session to 88% by the last. Essentially, the perception-only training was effective in improving perception of the two target vowels over the course of four sessions, which is reflected in the statistically significant improvements observed for the trained group on the Identification pre-to-post-test comparisons just discussed.

To determine the magnitude of the effect of perception-only training on perception, I utilized the pre- and post-test data from the Identification task with the synthetic shVp tokens. This testing condition most closely resembles the training condition, matching in task, stimuli quality, and phonetic context. This focus on the testing condition closest to training also follows the suggestions of Sakai and Moorman’s (submitted) meta-analysis. I calculated the Cohen’s $d$ within-group effect size and found the effect of the training was $d=1.52$. In order to interpret the magnitude of this effect, I consulted the findings of Sakai and Moorman (submitted), which report that the effects of perception-only training on perception for 21 unique samples ranges from $d=-2.14$ to $2.84$, with the median of 1.03. This leads to the interpretation that the perception-only training in the present experiment induced a large-sized effect on perception. A large-size magnitude is also consistent with the standards established by Plonsky and Oswald (2014) for the interpretation of effect sizes in the field of SLA in general.

In order to view the trained and the Control groups in light of the two baselines provided by the Bilingual Acquired and NS English groups, the bar graphs in Figure 6 can be visually compared to the dotted lines, which represent the two baseline groups. It is evident that the trained and Control groups do not reach the high scores of the Bilingual Acquired and NS English groups on any of the stimuli sets at pre- or post-tests. A one-way ANOVA of post-test Identification scores confirmed that the groups were different on the three different stimuli sets,
although Tukey’s posthoc comparisons revealed interesting results. On the shVp tokens, the Perception-Only trained group was significantly different from the Control group ($p=.001$) and the NS English group ($p=.01$), but they had reached the same level of performance of the Bilingual Acquired group ($p=.16$). On the synthetic fVt tokens, the trained group was the same as the Control ($p=.14$) and different from the Bilingual Acquired and NS English groups ($p=.01$ and $p=.00$, respectively). Finally, on the naturally recorded tokens, the trained group performed differently than all of the groups. They no longer performed similarly to the Control ($p=.03$), but had not quite reached the level of the Bilingual Acquired ($p=.001$) or the NS English group ($p=.00$). These results show that the perception-only training was effective in instigating improvement, so much so that the trained participants matched the level of performance of the Bilingual Acquired group on the synthetic shVp tokens, and they outperformed the Control group on the shVp and the naturally recorded tokens. However, the trained group did not reach the level of performance of the NS English group on any of the tasks.

The second perception measure, the Discrimination task, was used to determine if participants had sensitivity at a boundary between the /i/ and /ɪ/ phonemes regardless of their ability to identify the tokens in the context of lexical items. This type of analysis does not lend itself to inferential statistical comparisons; rather, the visual inspection of descriptive results on discrimination performance can help detect finer degrees of change that the Identification task may not have captured. The graphs in Figure 7 depict the spectral continuum for the Perception-Only and Control groups at pre-test (Graph A) and post-test (Graph B). The Bilingual Acquired and the NS English groups completed the Discrimination task only once, so their scores are repeated on both graphs as the dashed blue and grey lines, respectively, as points of comparison. The pre-test graph reveals that both the Perception-Only and Control groups indeed have a peak
of sensitivity at the Step 17-21 boundary, similar to the Bilingual Acquired and NS English groups, but the peaks are lower in amplitude. Additionally, the remaining step boundaries are more raised than the Bilingual Acquired and NS English groups. If the peak along a continuum is not substantially higher than the other step boundaries, the sensitivity at the category boundary is considered weaker. After training, the Perception-Only group developed a stronger peak at the Step 17-21 boundary, in fact stronger than either the NS English or Bilingual Acquired group; additionally all of the other boundaries decreased. These data can be interpreted as a sign that training effectively caused the Perception-Only group to acquire a spectral sensitivity that was nearly identical to the Bilingual Acquired and NS English groups. On the other hand, the Control group displayed a lowered sensitivity at the key boundary (Step 17-21), which could be signs of fatigue or lack of interest at the post-test session. Additionally, the Control group showed no peaks of sensitivity at any other point along the spectral steps in the post-test data.

Graphs A and B in Figure 8 depict the duration continuum at pre- and post-test. Neither the trained nor Control group showed a peak along this continuum. As there is no categorical boundary between /i/ and /ɪ/ based on duration, the fact that neither group shows a sensitivity spike along this continuum is a sign that both the trained and Control groups behave in a target-like manner in terms of duration both at pre- and post-tests. All groups have raised edges at the endpoints of the duration continuum. Braida, Lim, Berliner, Durlach, Rabinowitz, and Purks (1984) propose that these points of raised sensitivity are suggestive of a phenomenon called “perceptual anchoring”. One key component of this theory is that participants hear and interpret an auditory token in the Discrimination task in the context of all of the other tokens that they have already heard, and the auditory stimuli that are at the most extreme ends of the continuum act as anchor points against which all other tokens are compared. These anchors are perceived
Figure 7. Discrimination task scores across spectral steps for the Perception-Only (Per), Control (C), NS English (E), and Bilingual Acquired (B) groups. Graph (A) depicts pre-test scores, and (B) depicts post-test scores.
Figure 8. Discrimination task scores across duration steps for the Perception-Only (Per), Control (C), NS English (E), and Bilingual Acquired (B) groups. Graph (A) shows pre-test scores, and (B) shows post-test scores.
more accurately, resulting in higher sensitivity at these endpoints. The crucial factor of this theory is that the raised sensitivity at the endpoints is not indicative of a perceived categorical boundary, rather the endpoints are simply perceived differently because they are the most extreme versions of the auditory stimulus presented in the task. Thus, the raised endpoints in the present study’s Discrimination task results will not be interpreted to be an indication that the participants perceived category boundaries along the duration continua.

In sum, the perception data indicate that a two-hour perception-only training program effectively improved the trained groups’ perception of the English /i/ and /ɪ/ vowels. This result is typical of perception-only training, as the meta-analysis conducted by Sakai and Moorman found that perception-only training programs that were on average five hours in length had a medium-sized effect (d=.98) on participants’ perception. This experiment utilized a two-hour training and had a large-sized effect on trained participants’ perception (d=1.52). As it has been determined that perception-only training was effective on improving participants’ perception mode, now the crux of Experiment 1 can be addressed.

6.1.2. Research Question 2: Does perception-only training of two second language vowels improve the production of the same sounds?

The first analysis of participants’ productions is the NS judges’ identifications of extracted onsets and vowels. Table 11 presents the mean score of vowels correctly identified as /i/ and /ɪ/ by the three NS judges at pre- and post-test for all three tasks (i.e., Elicited Production, Passage Reading, and Picture Description) for the trained, Control, and baseline groups. It was expected that the Elicited Production task was the most likely to be susceptible to change after perception-only training because it elicited the most careful speech. The Passage Reading and Picture Description tasks elicited less careful speech and there was a smaller expectation that
these tasks would show change following perception-only training. Figure 9 shows a visual
depiction of the production task performance at pre- and post-test, in the bar graphs for the
trained and Control groups, and the dotted lines for the two baseline groups. I ran a 2 x 2 x 3
repeated measures ANOVA (Group by Time by Task) to determine if there was a change from
pre- to post-test on any of the tasks for either group. The assumption of normality was met, but
the assumption of homogeneity of variances was violated for the Elicited Production pre-test
data \((p=.002)\). I log-transformed the data, but this did not change the heterogeneity of variances.
Thus, the results from the Elicited Production data should be interpreted with caution.

Mauchley’s test of sphericity was not significant for Task \((p=.16)\) or Time by Task \((p=.53)\), thus
the assumption of sphericity was held. Results of the repeated measures ANOVA showed that
there was no significant Time by Group by Task interaction \((F(2, 58)=.43, p=.65, \eta^2_p=.02,\)
power=.12), nor was there a significant Time by Group interaction \((F(1, 29)=3.26, p=.08,\)
\(\eta^2_p=.10, \text{power}=.42)\). These results suggest that perception-only training did not have a
significant effect on participants’ productions as measured by the NS judges on any of the three
production tasks.

Although no significant results were found in the repeated measures ANOVA, I
calculated the Cohen’s \(d\) effect size with the Elicited Production data. This was the most
controlled production testing condition that elicited the most careful speech and thus the most
likely to show any change after training. The pre-to-post effect size was \(d=.29\), similar to but
smaller than the already small effect that Sakai and Moorman’s (submitted) meta-analysis found
for the production of vowels after perception-only training \((d=.46)\). Thus, with all of the data
thus far, it appears that the perception-only training of L2 vowels did not have an effect on
trained participants’ production of the vowels when utilizing the NS judgments of the Elicited
Table 11

*Mean (SD) native speaker identifications for three production tasks for the Perception-Only, Control, and baseline groups*

<table>
<thead>
<tr>
<th>Group</th>
<th>$n$</th>
<th>Elicited Production</th>
<th>Passage Reading</th>
<th>Picture Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pre</td>
<td>Post</td>
<td>Pre</td>
</tr>
<tr>
<td>Perception-Only</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>56.74</td>
<td>61.50</td>
<td>66.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(17.55)</td>
<td>(20.35)</td>
<td>(18.55)</td>
</tr>
<tr>
<td>Control</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>54.56</td>
<td>53.53</td>
<td>58.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(18.26)</td>
<td>(17.24)</td>
<td>(18.43)</td>
</tr>
<tr>
<td>NS English</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>98.80</td>
<td>--</td>
<td>95.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1.93)</td>
<td></td>
<td>(5.07)</td>
</tr>
<tr>
<td>Bilingual</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>94.86</td>
<td>--</td>
<td>86.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(9.10)</td>
<td></td>
<td>(13.05)</td>
</tr>
</tbody>
</table>

Production data.

In addition to the NS judges’ identifications, I also analyzed the production data via acoustical measurements. This type of analysis is similar to the Discrimination task analysis presented earlier for Research Question 1, in that it does not lend itself to inferential statistics but can help detect finer degrees of change that the NS judgments may not have captured. After extracting F1 and F2 values from all of the participants’ vowel productions, I normalized them into z-scores using the Bark Difference Metric so that values could be compared across individuals and groups. The graphs in Figure 10 and Figure 11 present all of the productions for the trained and Control groups at pre-test and again at post-test. These visualizations show that the productions of the /i/ and /ɪ/ vowels are largely overlapped for both groups; in other words, the vowels do not cluster as two distinct, separate categories. A close inspection of the Trained group’s post-test shows that the two categories may be starting to pull apart as many of the /ɪ/ productions are lowered and backed (i.e., moved to the right).
Figure 9. Production pre- and post-test scores for the Perception-Only (Per) and Control (C) groups on all three tasks. NS English (E) and Bilingual Acquired (B) group data are presented with dotted lines.
Figure 10. Z-score plots of the Trained group’s productions on all three tasks at pre- (A) and post-test (B). The red dots are /i/ productions, and the blue dots are /ɪ/ productions.
Figure 11. Z-score plots of the Control group’s productions from all three tasks at pre-test (A) and post-test (B). The red dots are /i/ productions, and the blue dots are /ɪ/ productions.
Table 12

Mean (SD) of percent overlap of /i/ and /ɪ/ for the Perception-Only, Control, and baseline groups at pre- and post-test

<table>
<thead>
<tr>
<th>Group</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perception-Only</td>
<td>52.32 (15.52)</td>
<td>46.90 (14.14)</td>
</tr>
<tr>
<td>Control</td>
<td>56.93 (10.45)</td>
<td>54.91 (13.45)</td>
</tr>
<tr>
<td>NS English</td>
<td>4.95 (6.37)</td>
<td>--</td>
</tr>
<tr>
<td>Bilingual Acquired</td>
<td>18.89 (19.36)</td>
<td>--</td>
</tr>
</tbody>
</table>

In order to investigate the productions quantitatively rather than relying on visual inspection alone, I calculated an overlap score for each individual’s two vowel productions using Johnson’s Overlap Simulator (Johnson, 2016), one overlap value for the pre-test and another for the post-test. This calculation measures the percentage of overlap of the two vowel clusters; in other words, one could imagine the measurement of the intersect portion of a Venn diagram figure. The mean percent of overlap for each group is presented in Table 12. The trained group’s overlap score is 52% at pre-test and 47% at post-test, while the Control group’s overlap score changes from 57% at pre-test to 55% at post-test. As points of comparison, the Bilingual Acquired group has a low overlap score of 19%, and the NS English group has an even smaller overlap of 5%. I ran a 2 x 2 repeated measures ANOVA (Group by Time) with the Perception-Only and Control group overlap data to test if there were any significant changes over time. The assumptions of normality and homogeneity of variances were held, and results showed that there was no significant Group by Time interaction ($F(1, 29)=.22, p=.65, \eta^2_p=.01, \text{power=.07}$). Although the trained group’s percentage of overlap for the two vowel categories does decrease, these data indicate that the perception-only training did not have a significant impact on the Perception-Only group’s productions of the two sounds.
One final way to inspect the acoustical data is to separate the produced tokens by task. Recall that the Elicited Production task elicits the most careful speech and thus is the most likely to show any change after training. The three graphs in Figures 12, 13, and 14 show the mean values for the /i/ and /ɪ/ vowels for each group at pre- and post-tests on the three different tasks. In general, it is clear that the trained and Control group productions for both /i/ and /ɪ/ sounds on the three tasks seem to cluster towards the prototypical English /i/ space, which is expected because this sound is the most similar to the Spanish /i/ vowel. However, Graph A shows that the Trained group’s production of /ɪ/ moves to a lower and more backed position at post-test, which brings it closer, but not fully towards the Bilingual Acquired and NS English /ɪ/ productions. This means that the participants who received perception-only training were able to improve their productions of English /ɪ/ to be more target-like on this one task, while the Control group’s /ɪ/ productions do not show this type of change at post-test. Although the more coarse-grained measure of participant productions (i.e., the NS judgments) could not capture the change in the trained group’s productions at post-test, an evaluation of the acoustic analysis of participants’ productions in the Elicited Production task shows quite a bit of improvement on the difficult to produce /ɪ/ vowel.

To summarize the results of Experiment 1, the perception-only training induced a significant, large-sized gain in perception for the participants who received the training, but the learning did not transfer to the production modality. However, acoustic analyses of the productions provided a finer-grained tool that showed the trained group’s production of the difficult vowel /ɪ/ underwent some change on one task, making it more target-like, but there was not clear evidence that participants produced /i/ and /ɪ/ as distinct categories at post-test.
Table 13

Z-score coordinates (Z3-Z2, Z3-Z1) for /i/ (A) and /ɪ/ (B) on the three production tasks at pre- and post-test

A

<table>
<thead>
<tr>
<th>Group</th>
<th>Elicited Production</th>
<th>Passage Reading</th>
<th>Picture Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Post</td>
<td>Pre</td>
</tr>
<tr>
<td>Perception-Only</td>
<td>1.48, 11.51</td>
<td>1.74, 11.41</td>
<td>1.78, 11.67</td>
</tr>
<tr>
<td>Control</td>
<td>1.55, 11.68</td>
<td>1.58, 11.64</td>
<td>1.56, 11.76</td>
</tr>
<tr>
<td>NS English</td>
<td>1.32, 11.99</td>
<td>--</td>
<td>1.25, 11.46</td>
</tr>
<tr>
<td>Bilingual, Acquired</td>
<td>1.52, 11.92</td>
<td>--</td>
<td>1.45, 11.34</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th>Group</th>
<th>Elicited Production</th>
<th>Passage Reading</th>
<th>Picture Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Post</td>
<td>Pre</td>
</tr>
<tr>
<td>Perception-Only</td>
<td>1.73, 11.41</td>
<td>1.86, 10.81</td>
<td>1.90, 11.45</td>
</tr>
<tr>
<td>Control</td>
<td>1.49, 11.46</td>
<td>1.56, 11.38</td>
<td>1.77, 11.50</td>
</tr>
<tr>
<td>NS English</td>
<td>2.70, 10.06</td>
<td>--</td>
<td>2.74, 10.50</td>
</tr>
<tr>
<td>Bilingual, Acquired</td>
<td>2.39, 10.27</td>
<td>--</td>
<td>2.40, 10.62</td>
</tr>
</tbody>
</table>

Figure 12. Bark normalized mean acoustic measurements for /i/ and /ɪ/ productions on the Elicited Production task for the Perception-Only (Per), Control (C), NS English, and Bilingual Acquired (B) groups.
Figure 13. Bark normalized mean acoustic measurements for /i/ and /ɪ/ productions on the Passage Reading task.

Figure 14. Bark normalized mean acoustic measurements for /ɨ/ and /ɪ/ productions on the Picture Description task.
6.2. Experiment 2

If Experiment 1 intended to focus on how the perception mode might influence the production mode, Experiment 2 sought to address the opposite question: Can the production mode influence the perception mode? To that end, three research questions were formally posed. The reporting in this section will follow the same plan as the previous one, although with the following differences. Readers should remember that in Experiment 2, two training groups participated, not one. Although both experimental groups completed the exact same training program, one group could hear their own voices (Production-Only) and the other could not (Production-Only, No Sound). In the present experiment, I will be reporting on the Production-Only and Production-Only, No Sound groups, in addition to presenting the same three comparison groups employed in Experiment 1: the Control, Bilingual Acquired, and NS English groups. For each analysis, these latter three groups’ data will be reproduced from the previous section for ease of comparison. First, I will address Research Question 1 by showing if either of the two production training regimens had any effect on production, as measured first by the NS judges. Next, I will report how the experimental groups performed during the four days of production-only training, in order to document any evidence of the learning process that may relate to the results observed in the production pre-post-test comparison. Subsequently, I will inspect the evidence for Research Question 1 from the perspective of acoustic analyses. Thereafter, I will address the second research question of whether production-only training influenced the perception mode. Lastly, I will turn to the third research question, which explicitly addresses a direct comparison of the two trained groups: Production-Only and Production-Only, No Sound. Namely, this third research question requires the overt comparison of the two trained groups with the intention of revealing the role played by the auditory feedback loop during
production-only training and whether access to the auditory feedback loop through self-produced auditory input (for the Production-Only group) versus no access to it (for the Production-Only, No Sound group) made any discernable difference in the learning of two L2 vowels.

**6.2.1. Research Question 1: Does production-only training of two second language vowels improve production of the same sounds?**

Table 14 presents the scores from the NS judges’ identifications of extracted onsets and vowels for both production-trained groups; the Control, Bilingual Acquired, and NS English group scores are presented as points of comparison. On the Elicited Production scores, the Production-Only group improved by 16%, and the Production-Only, No Sound group improved by 7%. On the remaining two tasks, the Production-Only group improved by 5% on the Passage Reading and decreased by 4% on the Picture Description task, and the Production-Only, No Sound group improved by 6% and 4% respectively. A 3 x 2 x 3 repeated measures ANOVA (Group by Time by Task) was used to determine if there was a difference between the two trained groups and the Control group on all of the tasks. All data were normally distributed and met the assumption of homogeneity, except the Elicited Production pre-test data. I log-transformed this data, but it did not improve normality or heterogeneity of variances, thus these data should be interpreted with caution. Mauchley’s test of sphericity was not violated for Time by Task ($p=.37$), but it was violated for Task ($p=.002$), so the Greenhouse-Geiser correction was used for any analysis with this variable. Results of the ANOVA showed that there was a significant Time by Task by Group interaction ($F(3.83, 80.39)=4.20, p=.004, \eta_p^2=.17$, power=.90). This means that any one of the three groups performed differently over time on one or some of the tasks, which is evidence of some learning. A series of paired samples $t$-tests revealed that the Production-Only group significantly improved on the Elicited Production task...
Table 14

Mean (SD) of native speaker identifications for three production tasks for the two experimental, one control, and two baseline groups

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Elicited Production</th>
<th></th>
<th>Passage Reading</th>
<th></th>
<th>Picture Description</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Post</td>
<td></td>
<td>Pre</td>
<td>Post</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Production-Only</td>
<td>14</td>
<td>48.22 (14.35)</td>
<td>63.96 (22.30)</td>
<td>53.06 (18.22)</td>
<td>58.83 (19.84)</td>
<td>61.71 (19.80)</td>
<td>57.19 (18.55)</td>
</tr>
<tr>
<td>Production-Only, No Sound</td>
<td>15</td>
<td>53.15 (14.98)</td>
<td>59.94 (22.02)</td>
<td>57.38 (18.20)</td>
<td>62.17 (17.92)</td>
<td>58.81 (19.24)</td>
<td>62.33 (18.91)</td>
</tr>
<tr>
<td>Control</td>
<td>16</td>
<td>54.52 (18.15)</td>
<td>53.80 (17.19)</td>
<td>57.94 (18.63)</td>
<td>58.70 (16.89)</td>
<td>62.86 (19.10)</td>
<td>60.74 (18.41)</td>
</tr>
<tr>
<td>NS English</td>
<td>20</td>
<td>98.80 (1.93)</td>
<td>--</td>
<td>95.50 (5.07)</td>
<td>--</td>
<td>94.37 (3.72)</td>
<td>--</td>
</tr>
<tr>
<td>Bilingual</td>
<td>16</td>
<td>93.80 (7.66)</td>
<td>--</td>
<td>85.94 (12.25)</td>
<td>--</td>
<td>87.28 (10.66)</td>
<td>--</td>
</tr>
</tbody>
</table>

Table 15

Mean percent accuracy score (SD) for two Production-Only training groups of two vowels on each day of training

<table>
<thead>
<tr>
<th>Group</th>
<th>Session 1</th>
<th>Session 2</th>
<th>Session 3</th>
<th>Session 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>i</td>
<td>i</td>
<td>i</td>
<td>i</td>
</tr>
<tr>
<td>Production-Only</td>
<td>32.86 (46.31)</td>
<td>13.89 (22.62)</td>
<td>82.30 (28.15)</td>
<td>56.10 (32.33)</td>
</tr>
<tr>
<td>Production-Only, No Sound</td>
<td>54.00 (37.76)</td>
<td>32.58 (27.65)</td>
<td>83.61 (15.82)</td>
<td>53.48 (31.77)</td>
</tr>
</tbody>
</table>
Figure 15. Production pre- and post-test scores for the Production-Only (Pro), Production-Only, No Sound (No Sound), and Control (C) groups on all three tasks. NS English (E) and Bilingual Acquired (B) group data are presented with dotted lines. A red star indicates a significant change from pre- to post-test.
(t(13)=−4.82, p<.00) and the Passage Reading task (t(13)=−2.47, p=.03), but not the Picture Description task (t(13)=1.60, p=.13). The Production-Only, No Sound group did not show any significant improvements on any of the three tasks: Elicited Production, t(14)=−1.83, p=.09; Passage Reading, t(14)=−1.89, p=.08; and the Picture Description, t(14)=−1.24, p=.24. The Control group also showed no significant improvement from pre- to post-test on any task: Elicited Production, t(15)=.26, p=.80; Passage Reading, t(15)=−2.8, p=.79; and the Picture Description, t(15)=.97, p=.35.

A visual inspection of the graphs in Figure 15 show that none of the trained or Control groups reached the level of performance of the Bilingual Acquired or the NS English groups. A one-way ANOVA of pre-test data from the two trained, one Control, Bilingual Acquired, and NS English groups showed that the groups performed differently on each task (Elicited Production, F(4,76)=174.46, p=.00; Passage Reading, F(4,76)=51.67, p=.00; and Picture Description, F(4,76)=40.80, p=.00). I next looked at the Games-Howell post-hoc analysis because equal variances were not assumed, which revealed that on the pre-test the two experimental groups were the same as each other and the Control group on all of the tasks except for the Elicited Production; in that case, the Production-Only group was significantly worse that the Production-Only, No Sound group (p=.04), but still the same as the Control group (p=.31). All three (experimental and Control) groups were also different at pre-test from the Bilingual Acquired and NS English groups on all tasks, at p=.00. At post-test, a one-way ANOVA of the five groups’ data showed that there were differences among the groups on all tasks (Elicited Production, F(4,76)=53.08, p=.00; Passage Reading, F(4,76)=52.63, p=.00; and Picture Description, F(4,76)=50.12, p=.00). Using the Games-Howell post-hoc analysis, it was revealed that the one Control and two experimental groups were still different from the Bilingual Acquired and NS
English groups on all tasks; none of the three had reached the high level performance of the baseline groups (all tasks and groups at \( p < .00 \)). Additionally, the post-hoc analysis showed that the one Control and two trained groups were not different from each other on any task (all tasks and groups at \( p > .05 \)). Thus, although the Production-Only group improved after training on the Elicited Production and the Passage Reading task, the group was still not different from the Control group on either of these tasks.

Inspecting the experimental group participants’ performance during training has the potential to shed light on any learning that occurred that was not captured in the production pre-and post-tests. Readers will recall that the production-only training utilized the Vowel Shapes program, in which participants’ voices were represented with a moving red dot and the target vowel was represented with a stable blue dot. Participants were instructed to manipulate their vowel productions (namely, the active articulators) so that their red dot reached the blue dot. Table 15 presents the percentage of the trials in which participants were able to reach the blue target. Note that the scores from Session 1 are low (14-54%). This occurred as a result of one or two reasons: (1) participants used the first training session to get acclimated to the Vowel Shapes program, and (2) this session was also used to determine each participant’s individual vowel target locations. To be more specific, some participants felt comfortable with the Vowel Shapes program immediately, while others took longer to understand how the program worked. Furthermore, I was able to determine some participants’ unique vowel target locations within a matter of seconds, while others required a lot of trial and error to locate their target vowel locations and to create a new blue target dot for them. If any time-consuming situation occurred on this first day of training, it was possible a participant was not able to take advantage of a full training session on Day 1, and in rare cases not complete more than five trials of Vowel Shapes
practice. (The effects of missing all or part of the first training session will be explored further in the Discussion chapter.) The variability of situations on training Day 1 led to a range of scores for this first session, falling anywhere between 0% and 100%. Thus, the low scores for Session 1 unfortunately do not necessarily reflect participants’ abilities on that day. In the end, I cannot reliably attribute the lower performance at Session 1 for the Production-Only group compared to the Production-Only, No Sound group to any difficulties the former had with either vowel. Rather, looking at the comparability of the scores on Day 2 offers a more trustworthy picture of participants’ abilities at the same point in time.

Over the course of the four-day training, but inspecting the process data starting with the more stable Session 2, participants in the Production-Only group improved their productions of the /i/ vowel modestly from 82% to 84% and the Production-Only, No Sound group improved this same vowel from 84% to 91%. These high scores reflect the ease participants’ felt in producing this vowel even before the training began, likely because this sound exists in Spanish. For the more difficult /ɪ/ vowel, by Session 4 the Production-Only group improved from 56% to 68%, and the Production-Only, No Sound group improved from 53% to 80%. The high standard deviations across both groups for both vowels throughout the four sessions are evidence that neither of the two versions of the production training helped all participants equally, nor did their performance for either training regimen converge by the last day of training. Under the conditions put forth in the training sessions under both the Production-Only condition and the Production-Only, No Sound condition, different trained participants acted and reacted differently, which induced different amounts of learning. For some, the training was very successful, and by the end of the program, they consistently scored 100% on both vowels. Others did not achieve
this success. The variability of the effectiveness of the training could have contributed to the low overall gains shown in the group means of the NS judgments’ of the participants’ productions.

One final note about the training scores is needed to properly interpret the percentages shown in Table 15. Recall that the perception training in Experiment 1 asked participants to listen to a sound and push a button to select which sound they thought they heard. If the participants could not distinguish the difficult /ɪ/ vowel, there would be evidence of random button pushing, and the resultant score would be close to 50%. Conversely, the production training necessarily involved a productive task. If a participant were not able to produce the correct vowel, the resultant score would be 0%. Thus a 50% score on the production training task is not the same as a 50% score on the identification training task. The fact that participants in the two trained groups in this experiment were correctly producing the difficult /ɪ/ vowel 68-80% of the time is quite a different feat than scoring the same percent on the identification training task.

Next, I calculated the magnitude of effect the production-only training had on participants’ productions of the target vowels. For this analysis, I used the NS judgments of the Elicited Production data because this task was closest to the training. The effect of production-only training for the Production-Only group was $d= .84$, and the effect of training on the Production-Only, No Sound group was a smaller effect of $d= .36$. To interpret these results in light of other pronunciation instruction effect sizes, I turn to the pronunciation meta-analysis conducted by Lee, Jang, and Plonsky (2015), a meta-analysis of the effects of pronunciation training programs across 110 studies. The researchers reported that the average effect size of pronunciation instruction when inspecting within-group contrasts (i.e., pre-to-post change) on a production test was $d= .83$. Lee et al. (2015) also reported moderating variables of the training programs, and for those pronunciation training programs lasting 4.25 hours or less (which they
consider short) the average effect size was $d=.62$. The average effect for pronunciation instruction that used a computer was $d=.75$, and the average effect of a training that focused on vowels was $d=.91$. Thus, in every regard, the Production-Only group from the present study benefitted from a very similar effect when compared to the meta-analysis effect sizes. By contrast, the Production-Only, No Sound group from this study had a smaller than average effect on participants’ productions. This is not unexpected since the Production-Only, No Sound group underwent a training that is quite different than real life, as participants were denied the benefit of hearing their own voices during pronunciation practice. Certainly, this type of training is less ecologically valid, except for in the case of deaf or hard-of-hearing L2 learners. Nevertheless, it is still interesting to note that there was a small effect of training.

The effect size calculation data is particularly interesting, not only because it shows that production of L2 vowels can improve after only four short training sessions, but also because of what it reveals about the effect of the auditory feedback loop. None of the participants in either production-only training group heard any other-produced exemplars of the target phonemes during training; the difference between the two groups lies in either having access to hearing their own voices during training or not having access to their own voices. This difference between training conditions proved to be important, as the Production-Only training group improved their productions at post-test while the Production-Only, No Sound group did not. The difference of the effects resulting from the auditory feedback loop of one’s own voice will be discussed further in the third research question of the present experiment.

The second series of analyses I conducted on the production data was through acoustic measurements. First, I normalized all of the tokens using the Bark Difference Metric. The $z$-scores from pre- and post-tests of the two trained groups are plotted on the vowel spaces in
At post-test, both trained groups seem to show the blue dots, or the /i/ tokens, beginning to pull away from the red dots, or the /ɨ/ tokens. It seems as though for both groups at post-test, the /i/ tokens are lowered and more backed, although there is still quite a substantial degree of overlap between the two vowels.

In order to examine the productions quantitatively rather than relying on visual inspection alone, I calculated each participant’s percentage of spectral overlap for their /i/ and /ɨ/ productions at pre- and post-test. Table 16 presents the overlap scores for both trained groups and the three comparison groups (i.e., Control, Bilingual Acquired, and NS English). A smaller overlap score at post-test would be evidence that the two vowel categories are diverging. At post-test, the Production-Only group’s spectral overlap score decreased by 14%, and the Production-Only, No Sound group’s score decreased by approximately 7%. In comparison, the Control group’s overlap score decreased by 2% at post-test. I ran a 3 x 2 repeated measures ANOVA (Group by Time) with data from the two trained groups and one Control group to see if there was a statistically significant effect of either type of training. The assumptions of normality and homogeneity of variances were met, thus no corrections to the data were needed. Results showed that there was not a Group by Time interaction ($F(2, 42)=1.29, p=.29, \eta^2_p=.06, \text{power}=.27$). Although both trained groups showed improvement from pre- to post-test, the change was not statistically reliable, nor different from each other or from the Control group.

For the final method of acoustical analysis, I divided the productions by task to discern any movement of the groups’ average productions from pre- to post-test. Figures 18-20 present three vowel spaces that correspond to the three production tasks (i.e., Elicited Production, Passage Reading, and Picture Description). There are two notable findings from the information displayed in these graphs. First, the pattern seems to be that the average /i/ and /ɨ/ tokens of all

103
Figure 16. Z-score plots of the Production-Only group’s productions on all three tasks at pre-test (A) and post-test (B). The red dots are /i/ productions, and the blue dots are /ɪ/ productions.
Figure 17. Z-score plots of the Production-Only, No Sound group’s productions on all three tasks at pre-test (A) and post-test (B). The red dots are /i/ productions, and the blue dots are /ɪ/ productions.
Table 16

Mean (SD) of percent overlap of the /i/ and /ɪ/ categories for all experimental, control, and baseline groups at pre- and post-test

<table>
<thead>
<tr>
<th>Group</th>
<th>Pre</th>
<th>Post</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production-Only</td>
<td>65.27 (8.63)</td>
<td>51.18 (18.22)</td>
</tr>
<tr>
<td>Production-Only, No Sound</td>
<td>58.33 (12.19)</td>
<td>51.56 (18.26)</td>
</tr>
<tr>
<td>Control</td>
<td>56.93 (10.45)</td>
<td>54.91 (13.45)</td>
</tr>
<tr>
<td>Bilingual Acquired</td>
<td>18.89 (19.36)</td>
<td>--</td>
</tr>
<tr>
<td>NS English</td>
<td>4.95 (6.37)</td>
<td>--</td>
</tr>
</tbody>
</table>

Table 17

Coordinates in z-scores (Z3-Z2, Z3-Z1) for /i/ (A) and /ɪ/ (B) in all three tasks at pre- and post-test

A

<table>
<thead>
<tr>
<th>Group</th>
<th>Elicited Production</th>
<th>Passage Reading</th>
<th>Picture Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Post</td>
<td>Pre</td>
</tr>
<tr>
<td>Production-Only</td>
<td>1.71, 11.39</td>
<td>1.58, 11.69</td>
<td>1.54, 11.30</td>
</tr>
<tr>
<td>Production-Only, No Sound</td>
<td>1.32, 11.63</td>
<td>1.54, 12.04</td>
<td>1.49, 11.48</td>
</tr>
<tr>
<td>Control</td>
<td>1.55, 11.68</td>
<td>1.58, 11.64</td>
<td>1.56, 11.76</td>
</tr>
<tr>
<td>NS English</td>
<td>1.32, 11.99</td>
<td>--</td>
<td>1.25, 11.46</td>
</tr>
<tr>
<td>Bilingual, Acquired</td>
<td>1.52, 11.92</td>
<td>--</td>
<td>1.45, 11.34</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th>Group</th>
<th>Elicited Production</th>
<th>Passage Reading</th>
<th>Picture Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Post</td>
<td>Pre</td>
</tr>
<tr>
<td>Production-Only</td>
<td>1.64, 11.39</td>
<td>1.92, 10.87</td>
<td>1.70, 11.18</td>
</tr>
<tr>
<td>Production-Only, No Sound</td>
<td>1.34, 11.40</td>
<td>1.58, 11.33</td>
<td>1.65, 11.30</td>
</tr>
<tr>
<td>Control</td>
<td>1.49, 11.46</td>
<td>1.56, 11.38</td>
<td>1.77, 11.50</td>
</tr>
<tr>
<td>NS English</td>
<td>2.70, 10.06</td>
<td>--</td>
<td>2.74, 10.50</td>
</tr>
<tr>
<td>Bilingual, Acquired</td>
<td>2.39, 10.27</td>
<td>--</td>
<td>2.40, 10.62</td>
</tr>
</tbody>
</table>
experimental and Control groups cluster around the prototypical English /i/ area. Again, this is not surprising as Spanish has one vowel, /i/, in this portion of the vowel space. The second valuable piece of information gathered from Figure 18 and Figure 20 is that both the Production-Only group and the Production-Only, No Sound group show movement of /i/ and /ɪ/ vowels in a target-like direction at post-test. In Figure 18, representing the Elicited Production task, the Production-Only group begins at pre-test with both the /i/ and /ɪ/ averages at nearly the same location. Then at post-test, both trained groups’ /i/ vowels raise, and both trained groups’ /ɪ/ vowels lower and move back, with the Production-Only group’s /ɪ/ at post-test moving dramatically toward the target area. In Figure 20, representing the Picture Description task, the Production-Only group again shows backed movement in the post-test for the /ɪ/ average. This is unexpected given that this is the task with the least monitored speech and therefore was expected to be the most resistant to change. Yet, it shows evidence of improvement for the difficult /ɪ/ vowel for one of the trained groups. Overall, these graphs are able to show that both trained groups show some improvement in the production of the two vowels, and that the Production-Only Trained group’s improvement is stronger and more dramatic.

In sum, the two conditions of the production-only training program seem to be different in terms of how much change they effected: The production-only training with access to hearing the auditory feedback loop was successful in improving participants’ productions, while the production-only training without access to hearing the auditory feedback loop was not. The Production-Only group improved in production to a medium and statistically robust degree on the Elicited Production task, and the Production-Only, No Sound group showed a small improvement that was not statistically reliable. The acoustical analyses were able to show finer
Figure 18. Bark normalized mean /i/ and /ɪ/ productions for the control and two trained groups on the Elicited Production task.

Figure 19. Bark normalized mean /i/ and /ɪ/ productions for the control and two trained groups on the Passage Reading task.
degrees of change in both trained groups productions’ on the various tasks, with the Production-Only group consistently showing greater change.

Since I have analyzed the two types of production-only trainings’ effects on production, I will now move on to the core of Experiment 2 and analyze the participants’ perception at pre- and post-test in order to ascertain whether production-only training can affect perception.

6.2.2. Research Question 2: Does production-only training improve the perception of two target second language vowels?

Table 18 presents the Identification task data from the two experimental, one Control, and two baseline groups. By post-test, the Production-Only group improved their perception by 16% on the shVp tokens, 9% on the fVt tokens, and 11% on the naturally recorded tokens. The Production-Only, No Sound group improved 19%, 28%, and 4% on the same three tasks.
Table 18
Mean (SD) for three sets of stimuli on the Identification task at pre- and post-test

<table>
<thead>
<tr>
<th>Group</th>
<th>shVp</th>
<th>fVt</th>
<th>Natural Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Post</td>
<td>Pre</td>
</tr>
<tr>
<td>Production-Only</td>
<td>49.70</td>
<td>66.07</td>
<td>62.50</td>
</tr>
<tr>
<td>Production-Only, No Sound</td>
<td>49.39</td>
<td>68.45</td>
<td>45.33</td>
</tr>
<tr>
<td></td>
<td>(18.87)</td>
<td>(16.48)</td>
<td>(20.70)</td>
</tr>
<tr>
<td>Control</td>
<td>44.69</td>
<td>52.67</td>
<td>50.47</td>
</tr>
<tr>
<td></td>
<td>(13.96)</td>
<td>(17.82)</td>
<td>(19.31)</td>
</tr>
<tr>
<td>NS English</td>
<td>85.88</td>
<td>--</td>
<td>88.63</td>
</tr>
<tr>
<td></td>
<td>(5.98)</td>
<td>(5.32)</td>
<td>(1.92)</td>
</tr>
<tr>
<td>Bilingual</td>
<td>81.72</td>
<td>--</td>
<td>83.59</td>
</tr>
<tr>
<td></td>
<td>(6.17)</td>
<td>(8.74)</td>
<td>(2.25)</td>
</tr>
</tbody>
</table>

Figure 21 shows a visual depiction of the Identification task performance at pre- and post-test, in the bar graphs for the two trained groups and the one Control group, and in the dotted lines for the two baseline groups. I ran a 2 x 3 x 3 repeated measures ANOVA (Time by Group by Stimuli) with just the two experimental and one Control groups’ data to test if there was a significant difference between groups from pre- to post-test. The assumption of homogeneity was not violated, but the assumption of normality was violated for the pre-test data of the naturally recorded stimuli, thus I log transformed the naturally recorded pre- and post-test data. Mauchley’s test of sphericity was not violated for Stimuli (p=.09) or Time by Stimuli (p=.48), so the assumption of sphericity was held. Results showed that there was a significant Time by Group by Stimuli interaction ($F(4, 84)=3.47, p=.01, \eta_p^2=.14, \text{power}=.84$), which means that the groups performed differently on the various sets of stimuli at pre- and post-tests. There was also a Time by Group interaction ($F(2, 42)=4.21, p=.02, \eta_p^2=.17, \text{power}=.71$). A series of paired samples $t$-tests revealed that the Production-Only group improved from pre- to post-tests on all
Figure 21. Identification task pre- and post-test scores for the two experimental one control group on each of the three sets of stimuli. NS English (E) and Bilingual Acquired (B) group data are represented by the dotted lines. A red star indicates a significant difference at post-test.
stimuli sets (shVp, \( t(13)= -2.91, p = .01 \); fVt, \( t(13)= -2.63, p = .02 \); and naturally recorded, \( t(13)= -2.32, p = .04 \)). The Production-Only, No Sound group improved on the shVp stimuli (\( t(14)= -2.72, p = .02 \)) and the fVt stimuli (\( t(14)= -3.94, p = .001 \)), but not the naturally recorded stimuli (\( t(14)= -1.60, p = .13 \)). The following parallels the reporting in Experiment 1: The Control group showed significant improvement from pre- to post-test on the shVp stimuli (\( t(15)= -2.41, p = .03 \)), but not the fVt (\( t(15)= -1.77, p = .10 \)) or the naturally recorded stimuli (\( t(15)= 2.01, p = .06 \)).

Although the two trained groups improved after production-only training, a visual inspection of the bar graphs and the dotted lines clearly show that the two trained groups do not quite reach the perception levels of the Bilingual Acquired and the NS English group scores on any set of stimuli. A one-way ANOVA of post-test Identification scores shows that the groups performed differently on the three different sets of stimuli. A Tukey’s post-hoc comparison revealed that the two trained groups performed the same as each other on all three sets of stimuli (shVp, \( p = .99 \); fVt, \( p = 1.00 \); naturally recorded, \( p = 1.00 \)). Additionally, the Production-Only, No Sound group reached the performance level of the Bilingual Acquired group on the shVp stimuli (\( p = .06 \); both trained groups reached the Bilingual Acquired group’s performance level of the fVt stimuli (Production-Only, \( p = .19 \); Production-Only, No Sound, \( p = .25 \)); and neither of the trained groups reached the Bilingual Acquired performance level on the naturally recorded stimuli (Production-Only, \( p = .004 \); Production-Only, No Sound, \( p = .002 \)). Neither of the trained groups reached the level of NS English performance on any of the sets of stimuli (\( p < .03 \) on all stimuli sets). Finally, the two trained groups outperformed the Control on all sets of stimuli (\( p < .02 \)), except in one case, where the Production-Only group performed similarly to the Control group on the shVp stimuli (\( p = .06 \)). In sum, the production-only training helped the trained groups reach
the level of Bilingual Acquired performance on the Identification task on some, but not all, of the stimuli sets.

Next, I calculated the effect sizes of the production-only training on the perception scores, using the Identification task shVp stimuli because it was the most like the training as possible. The effect of production-only training on the shVp identification for the Production-Only group was $d=0.98$ and for the Production-Only, No Sound group was $d=1.08$. There are no previous studies that have attempted to reach the perception mode through strictly isolated production-only training, so I cannot compare these effects to prior published research. Nevertheless, using the range of Cohen’s $d$ scores from Sakai and Moorman (submitted), Lee et al. (2015), and Plonsky and Oswald (2014), the effect sizes of this present study can be said to be of a medium effect. Thus, the production-only training had a medium-sized effect on the perception mode, regardless of the difference between two training regimens. Thus, it is interesting that tapping access to the auditory feedback loop (for the Production-Only group) versus having no access to it (for the Production-Only, No Sound group) made no discernable difference on cross-modal learning from the production to the perception mode. It is also particularly interesting that the boost the Production-Only group received on the production outcome for the Elicited Production task did not also afford them a boost in the perception mode.

Finally, I will report the effects of production-only training on participants’ discrimination scores. The graphs in Figure 22 show that the two trained groups have quite similar discrimination scores at pre-test, and both have a peak at the target step boundary. At post-test, both groups increase the intensity of the peak at the target boundary, while the Production-Only group decreases sensitivity at the other boundaries and the Production-Only, No Sound group maintains the pre-test levels of sensitivity at other boundaries. In this sense, the
Figure 22. Discrimination task scores across spectral steps for the Production-Only (Pro), Production-Only, No Sound (Pro, NS), Control (C), NS English (E), and Bilingual Acquired (B) groups at pre-test (Graph A) and post-test (Graph B).
Figure 23. Discrimination task scores across duration steps for the Production-Only (Pro), Production-Only, No Sound (Pro, NS), Control (C), NS English (E), and Bilingual Acquired (B) groups at pre-test (Graph A) and post-test (Graph B).
Production-Only group aligns their spectral sensitivity closer to target-like discrimination than the Production-Only, No Sound group. As far as the participants’ discrimination along the duration continuum, shown in Figure 23, the Production-Only group has a peak at Step 3-5 at pre-test but, by post-test, this group decreases sensitivity at that boundary and only shows raised endpoints. The Production-Only, No Sound group begins with a relatively flat line along the duration continuum at pre-test and maintains the flat line at post-test, with the exception of a raised endpoint at the upper end of the continuum. As previously reported, raised endpoints along the duration continuum can be interpreted as effects of perceptual anchoring and are not indicative of categorical boundaries (Braida et al., 1984).

To summarize the research questions of Experiment 2 until this point, the production-only training afforded the Production-Only group improvements in the production mode, and the production-only training had positive effects on both trained groups’ perception. The Production-Only group had a consistent and reliable edge in production gains over the Production-Only, No Sound training group. In the Elicited Production task with NS judges, the Production-Only group improved their productions to a medium-sized, statistically reliable effect, while the Production-Only, No Sound group improved on this task to a small-sized and statistically not reliable effect. Both groups improved their overlap score, with the Production-Only group improving more, and both groups improved their average /i/ and /ɪ/ productions in terms of acoustic cues, again, with the Production-Only group showing greater, more target-like changes. In terms of perception, both of the two trained groups improved their perception on the Identification task with the shVp stimuli to a medium and statistically robust degree. That perception improved for both groups is the first evidence in this domain of literature that the perception mode can be accessed through and positively influenced by the production mode.
Because it is valuable to see how all three trained groups (i.e., Perception-Only, Production-Only, and Production-Only, No Sound) performed on the pre- and post-test tasks in comparison to each other, I have included tables in Appendix H that compile all of the data for all of the pre- and post-tests for the three trained, one Control, and two baseline groups. I am providing this combined version of the data for ease of comparing scores across experiments.

6.2.3. Research Question 3: Does access to the auditory feedback loop (for the Production-Only group) versus no access (for the Production-Only, No Sound group) influence the development of second language speech perception and production after production-only training?

An element of the data that has been present throughout the reporting of Experiment 2 is the implicit comparison of the two trained groups: Production-Only and Production-Only, No Sound. This third research question allows for the overt comparison of the two groups with the intention of revealing the effects of self-produced auditory input, or in other words, the auditory feedback loop, during production-only training.

Table 19 presents a summative side-by-side comparison of the two groups’ performance on all of the perception and production tasks at pre- and post-test. A reading of the data presented in this table shows that the Production-Only group generally has larger gains on the production tasks, and the Production-Only, No Sound group generally has larger gains on the perception tasks. But a closer inspection of the perception data reveals that although the No Sound group has larger gains on the Identification task with the shVp and fVt stimuli sets, the two groups’ scores converge at post-test on all three stimuli sets. To explain this, I will present individual participants’ scores, which had a large impact on the Production-Only, No Sound group averages.
Table 19

A comparison of the Production-Only and Production-Only, No Sound training groups’ performance on perception and production tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Production-Only</th>
<th></th>
<th>Production-Only, No Sound</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-test</td>
<td>Post-test</td>
<td>Gain Score</td>
<td>Pre-test</td>
</tr>
<tr>
<td>Perception: Identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shVp</td>
<td>49.70</td>
<td>66.07</td>
<td>16.37</td>
<td>49.39</td>
</tr>
<tr>
<td>fVt</td>
<td>62.50</td>
<td>72.02</td>
<td>9.52</td>
<td>45.33</td>
</tr>
<tr>
<td>Naturally Recorded</td>
<td>68.16</td>
<td>79.17</td>
<td>11.01</td>
<td>73.61</td>
</tr>
<tr>
<td>Production</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS Judges’ Identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elicited Production</td>
<td>48.22</td>
<td>63.96</td>
<td>15.74</td>
<td>53.15</td>
</tr>
<tr>
<td>Passage Reading</td>
<td>53.06</td>
<td>58.83</td>
<td>5.77</td>
<td>57.38</td>
</tr>
<tr>
<td>Picture Description</td>
<td>61.71</td>
<td>57.19</td>
<td>-4.52</td>
<td>58.81</td>
</tr>
<tr>
<td>Spectral Overlap Score</td>
<td>65.27</td>
<td>51.18</td>
<td>14.09</td>
<td>58.33</td>
</tr>
</tbody>
</table>

Note. All gain scores (shown in shaded columns) were calculated by subtracting the pre-test from the post-test, one exception being the Spectral Overlap Score. On this type of analysis, a lower score is better, thus the post-test was subtracted from the pre-test to derive the gain score.

The accuracy scoring of the Identification task depends on a dichotomous choice, meaning that a score of 50% is indicative of no knowledge, or random guessing. Conversely a lower numerical score, for example 20%, signifies an ability to correctly perceive and identify the two sounds, but with a transposed lexical association. Thus, a score of 20% and a score of 80% indicate the same degree of ability to identify the vowel sound, but with incorrect lexical knowledge. At the pre-test stage, two participants (#210 and #212) out of fourteen in the Production-Only group scored below 30% on the shVp stimuli, and no participants scored lower than 30% on the fVt stimuli. On the other hand, four participants (#207, #225, #298, and #607) out of fifteen in the Production-Only, No Sound group scored below 30% on the shVp stimuli set and also on the fVt stimuli set (Participant #207 scored 24% on the shVp stimuli and 33% on the fVt stimuli). Because four participants in the Production-Only, No Sound group scored very low on the pre-test of both of these stimuli sets, the overall mean scores were lowered. After training, all six of these participants in the Production-Only and Production-Only, No Sound groups had
fixed the lexical transposition problem, and most scored over 80% accuracy on the Identification shVp and fVt post-tests. In comparison, the Control group had four participants who scored less than 30% at pre-test on one of the synthetic stimuli sets, and only one of these participants had fixed the problem by post-test. Thus, this transposition phenomenon seems to be reliably fixed through training, and it affords the group an overall large jump in mean score at post-test. It follows that the larger gain score seen in the Production-Only, No Sound group was influenced by the number of participants who had a transposition problem at pre-test, which makes their gain score seem inflated in comparison to the Production-Only group.

After examining individual participants’ scores, it is evident that the Production-Only and Production-Only, No Sound groups’ gain scores on the Identification task measures seem more comparable. The two groups’ effect size scores on the Identification shVp stimuli were similar at $d=.98$ and $d=1.08$, and the small benefit afforded to the Production-Only, No Sound group is a wash, given the information about the participants who had a transposition problem in this group. Thus, for perception, the presence or absence of the auditory feedback loop does not seem to affect the medium and statistically reliable perception gains afforded to participants after production-only training.

For production, on the other hand, the picture is different. The Production-Only group showed a regular advantage in production measures over the No Sound group. On the Elicited Production task using the NS judges’ identification analysis, the Production-Only group improved by 16% while the No Sound group improved by 7%. On this same task using the acoustical analysis, the Production-Only group made larger improvements moving the average /ɪ/ production towards the Bilingual Acquired and NS English productions. Finally with the Spectral Overlap score, the Production-Only group decreased the two vowels’ overlap by 14%
while the No Sound group only decreased it by half that amount. Thus, with regard to production, the auditory feedback loop seems to have provided an advantage for the Production-Only group that the No Sound group did not receive.

To understand this picture better, it is necessary to also look at the two groups’ performance during the production-only training using the Vowel Shapes program (refer back to Table 15 on page 96). By Session 4, the No Sound group was able to reach the /i/ target in 91% of the trials and the /ɪ/ target 80% of the time, compared to members of the Production-Only group who were able to reach the /i/ target 84% of the time and the /ɪ/ target 68% of the time. By this measure, it seems like the No Sound group was on its way to showing greater improvement in the post-tests, as they had mastered the Vowel Shapes training program better than the Production-Only group. But this trajectory did not materialize. Better success on the training under conditions of No Sound did not translate to more improvements for this group on the production post-test measures, where performance was again in a more natural environment, with access to self-produced auditory input. In this case, access to the auditory feedback loop gave a boost in production post-test measures, but not in the actual production-only training.

To summarize the entirety of Experiment 2, production-only training that offered participants no other-produced auditory examples of the target phonemes was successful in improving the Production-Only group’s productions on the Elicited Production task to a medium-sized degree; this group also improved their perceptions on the Identification task with the synthetic shVp tokens to a significant and medium-sized degree. The production-only training that offered participants no other- or self-produced auditory examples of the target phonemes (Production-Only, No Sound condition) was not successful in improving their production scores on the Elicited Production task, but participants who underwent this condition
of training did significantly improve their perception scores to a medium-sized degree on the Identification task with synthetic shVp tokens, just as the participants in the more natural Production-Only condition did. Thus, crucially, the production-only training transferred learning to the perception mode. There are three important conclusions to be drawn from the comparison of the two production training regimens: (1) deprivation of the auditory feedback loop provided an advantage during the production-only training that did not translate to an advantage on the production post-test measures; (2), access to the auditory feedback loop provided an advantage in the production post-test measures; and (3) presence or absence of the auditory feedback loop did not influence the perception testing outcomes.

6.3. Experiment 3

6.3.1. Research Question 1: How comparable and useful are bilingual and native speaker baseline groups?

One objective of collecting data from twenty NSs of English and sixteen Spanish-English bilinguals in the present study was to employ both of these groups as baselines against which I could interpret the performances of the experimental and control group participants in the perception-only training of Experiment 1 and the production-only training in Experiment 2. However, it is worth asking the question of just how similar or different these two baseline groups perform on perception and production tasks targeting the English vowels /i/ and /ɪ/, with the ultimate goal being to add information about ideal and useful baselines for use in research that focuses on effectiveness of phonological training specifically, and on L2 phonological development and ultimate attainment more generally. As a reminder (see the Method Chapter), the NS English group was comprised of functionally monolingual English users who had a predominantly monolingual upbringing and did not use a second language to an extensive level
in their daily adult lives. The Spanish-English bilinguals that comprised the Bilingual Acquired baseline were L2 participants who achieved over 80% on both the perception and production tasks at the prescreening stage. Being so close to ceiling on the pre-test measures, speech training on the /i/ and /ɪ/ phonetic distinction was deemed unnecessary for them, thus these participants were withheld from continuing with the experimental portion of the dissertation. As an additional reminder, in all of the analyses reported in Experiment 1 and 2, when the Bilingual Acquired and NS English groups were submitted to statistical tests, they were never found to be different from each other ($p > .05$ on all perception and production measures).

Beginning with the perception tasks, Figure 24 displays three graphs that present the range of participant scores on the Identification task with three different types of stimuli. The blue lines represent each Bilingual Acquired participant, the grey band is the range of NS English scores, and the grey arrow at the side of the graph points to the mean NS English score. First investigating the synthetic shVp tokens, it is apparent that the range of performance for both groups was quite broad. At the same time, it is also immediately evident that only three Bilingual Acquired participants performed outside of the range of NS English scores (Participants #408, #237, #241), and at that, the scores are only slightly below the NS range. On the Identification task with the synthetic fVt stimuli, the same pattern occurs, with a broad range of scores for both groups, but this time five Bilingual Acquired participants are below the NS English range. Yet, it is interesting that the lowest performer in the shVp stimuli set (Participant #408) was the highest performer on the fVt set of stimuli. Similarly, Participant #237, who scored low on the synthesized shVp tokens, had 90% accuracy on the synthesized fVt tokens. This is an indication that low performance on one stimuli set may not necessarily follow low performance on another set. Graph C, which presents the data from the naturally recorded tokens, shows a different
Figure 24. NS English and Bilingual Acquired scores on the Identification task with three types of stimuli: (A) shVp, (B) fVt, and (C) naturally recorded tokens. One blue line represents one Bilingual Acquired participant, the grey band represents the range of NS English scores, and the grey arrow indicates the NS English group’s mean score.
picture than the synthesized stimuli pattern. Now, all of the participants in both groups converge at ceiling. All of the participants in both groups scored above 90%, which is evidence that participants did not have trouble perceiving words that are most similar to natural speech. To summarize the Identification task performance, most Bilingual Acquired participants scored within native speaker ranges on all stimuli. The naturally recorded tokens, which were most similar to everyday speech, was the easiest for the Bilingual Acquired and NS English groups alike, as they both scored very high. Perception was more difficult for both groups on the synthesized tokens, but the bilinguals showed evidence of more difficulty, in that their scores were affected more than the native speaker group.

Next, I analyzed the results of the Discrimination task, which would reveal if participants had sensitivity at a categorical boundary between the /i/ and /ɪ/ vowels along two continuums. Graph A in Figure 25 shows the Bilingual Acquired and NS English groups’ sensitivity along the spectral continuum, and it is clear that the two groups are nearly identical at each boundary step. Both groups have a low sensitivity across the continuum and show only one spike that peaks at Step 17-21, which is an indication that participants perceived the spectral boundary of the two phoneme categories between these two steps. The second graph, B, displays the duration continuum. Again, the groups are quite similar in their perceptions across this continuum as well. Because there is no categorical boundary of /i/ and /ɪ/ determined by duration in English, the line across these steps should be somewhat even. The Bilingual Acquired group’s relatively flat line is a sign that they ignored the duration cue when listening for a distinction between /i/ and /ɪ/. On the other hand, the NS English group shows a small peak at the short end of the continuum (Steps 1-3). As reported earlier in this chapter, an increase at an endpoint is not likely a categorical boundary, but rather the result of a psychophysical anchoring effect associated with
Figure 25. Bilingual Acquired and NS English group performance on the Discrimination task. Graph A displays sensitivity along the spectral continuum, and Graph B displays sensitivity along the duration continuum.
continuum endpoints (Braida et al., 1984). Overall, the two groups performed quite similarly on discriminating spectral and duration cues of the two English vowels.

In terms of production, the Bilingual Acquired group shows more variation on these tasks than the perception tasks, and scores were susceptible to falling below the NS English range more often than the perception tasks. Figure 26 depicts the NS judges’ identifications of participants’ productions on three tasks: Graph A displays the Elicited Production scores, Graph B, the Passage Reading, and Graph C, the Picture Description. The NS English group scored very high on the Elicited Production task with the average score being 98.80% and the standard deviation 1.93. All of the NS English participants scored at or above 93%. Similarly, all Bilingual Acquired participants scored at or above 93% on this task, with the exception of three participants: Participants #274, #237, and #255 (listed from lowest to highest score). Participant #237 also scored outside of the NS English range on the Identification shVp stimuli, and Participant #255 also scored outside of the NS English range on the Identification fVt stimuli. On the other hand, although Participant #274 scored low on the Elicited Production task, she was a very good perceiver of the two vowels and scored 93% on the Identification task with shVp stimuli, 87% on the fVt stimuli, and 96% on the naturally recorded stimuli. Again, there is a pattern that a lower score on one task does not follow lower scores on another, even across the modalities.

As the production tasks elicit less careful speech, the variation in the Bilingual Acquired group’s performance increases. In the Passage Reading task, more Bilingual Acquired participants score outside of the NS English range (n=5) than the Elicited Production task (n=3). But the most variation in the Bilingual Acquired group’s scores is in the Picture Description task, which elicits the least careful speech of all three production tasks. In this case, nearly half
Figure 26. NS English and Bilingual Acquired groups' scores on three production tasks: (A) Elicited Production, (B) Passage Reading, and (C) Picture Description. One blue line represents one Bilingual Acquired participant, the grey band represents the range of NS English scores, and the grey arrow indicates the NS English group’s mean score.
of the Bilingual Acquired group \((n=7)\) scored outside of the NS English range. With these production data, it seems that the Bilingual Acquired participants produce less easily identifiable (by NS judges) \(/i/\) and \(/ɪ/\) tokens as the speech elicitation method yields less and less careful speech.

In order to see a clearer picture of the low performing Bilingual Acquired participants on both the perception and production tasks, I created a table that lists which participants scored outside of the NS English ranges. Table 20 reveals the ten bilingual participants, out of sixteen, who fell outside of the NS English range on at least one task. Four only performed at this lower level on one task, either on the more difficult synthesized tokens in the Identification task or the more difficult production task that elicited the most unmonitored speech, the Picture Description task. The remaining six participants (i.e., 38% of the entire Bilingual Acquired group) scored outside of the NS English range on two or more tasks; the entirety of these participants scored outside of the NS English range on the Picture Description task. The table makes it obvious that it is not the case that the same participants continue to be the low scorers

Table 20

Bilingual Acquired participants who scored lower than the NS English range on the perception tasks or production tasks

<table>
<thead>
<tr>
<th>Participant #</th>
<th>shVp</th>
<th>fVt</th>
<th>Naturally Recorded</th>
<th>Elicited Production</th>
<th>Passage Reading</th>
<th>Picture Description</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>223</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>228</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>259</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>408</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>242</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>221</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>274</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>237</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>241</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>255</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

Note: The “Sum” column presents the total number of tasks in which the participant scored outside of the NS English range, and the shaded rows highlight participants who performed outside of the NS English range on only one task.
on all of the tasks; rather, some participants score lower on some tasks and within the NS English range on the others.

In order to compare the two groups’ productions through acoustical analyses, I normalized the formant values using the Bark Difference Metric. The scatter plots in Figure 27 show all of the Bilingual Acquired productions in Graph A and all of the NS English productions in Graph B. Although both groups show clear clustering of the red (/i/) and blue (/ɪ/) dots, the Spectral Overlap analysis revealed that the Bilingual Acquired group had an average of 19% overlap of the two vowel categories, and the NS English group had an average overlap of only 5%.

Rather than displaying productions across the three tasks compiled together like in Figure 27, the graphs in Figure 30 display both groups’ average /i/ and average /ɪ/ now plotted on three graphs separated by task. Figure 28 depicts the mean productions from the Elicited Production task, Figure 29, the Passage Reading task, and Figure 30, the Picture Description task. With these illustrations, it is again apparent that both groups produced the two vowels as distinct categorical sounds; however, the Bilingual Acquired group consistently produced the two vowels closer together in the vowel space, while the NS English group unfailingly produced the two categories further apart on the vowel space. In other words, the Bilingual Acquired group produced the two vowels less distinctly than the NS English group. That the productions were acoustically closer together in the vowel space could be one cause for the NS judges’ misidentification of tokens from the previously discussed method of analysis of productions. A final interesting observation from these three vowel plots is the movement of both groups’ mean /i/ and /ɪ/ vowels as the production elicitation method moves from more to
Figure 27. Graph A depicts the Bilingual Acquired group’s /i/ and /ɪ/ vowels and B presents the vowels for the NS English group. The /i/ vowel is represented by red dots, and the /ɪ/ vowel by navy blue dots.
Figure 28. Average /i/ and /ɪ/ vowel productions by NS English and Bilingual Acquired groups on the Elicited Production task.
Figure 29. Average /ɪ/ and /iː/ vowel productions by NS English and Bilingual Acquired groups on the Passage Reading task.

Figure 30. Average /ɪ/ and /iː/ vowel productions by NS English and Bilingual Acquired groups on the Picture Description task.
less careful speech in the production tasks. In the most careful speech, the Elicited Production task, both groups produce the most extreme version of the two vowels of any of the tasks. Then, in the Passage Reading task, which is slightly less monitored speech, both groups’ two vowels move closer together; both groups’ /i/ tokens are markedly lowered. Finally in the Picture Description task, which elicits the most free, unmonitored speech of all the tasks, the average /i/ and /ɪ/ productions of both groups are the closest in distance, and both groups’ /ɪ/ tokens are notably more fronted than in the other tasks.

Overall, it is clear that the Bilingual Acquired group had two distinct categories for the /i/ and /ɪ/ vowels. These bilingual participants, who were sampled into this group because they were able to score at 80% or higher accuracy on all pre-tests, had achieved categorical distinction between /i/ and /ɪ/ in perception and production across the board, with the exception of small vulnerabilities of performance in the identification of synthetic tokens and the more unmonitored speech of the production tasks. The Bilingual Acquired group’s Discrimination task sensitivities along the spectral steps were strong evidence that this group had a definitive categorical boundary between the /i/ and /ɪ/ vowels. In isolated tokens of naturally produced speech, the Bilingual Acquired group was able to identify the /i/ and /ɪ/ values extremely well, nearly at the same level as the NS English group. In the more difficult Identification task of manipulated, synthetic tokens, there was more evidence of variation in both bilingual and native speaker groups; however some participants from the Bilingual Acquired group seemed to be more susceptible to declining performance than the NS English group. In production, both groups produced clearly distinct /i/ and /ɪ/ vowels, although the Bilingual Acquired group’s productions were closer on the spectral continuum than the NS English group’s vowels. Finally,
some bilingual participants who exhibited difficulty on one task, did not necessarily have difficulty on any other.

In the next chapter, I will discuss the results of the research questions from all three experiments of this dissertation, paying particular attention to the implication of the results on theoretical models of second language speech acquisition, cognition, L2 training methodology, and L2 pedagogy.
Chapter 7: Discussion and Conclusion

This dissertation attempted to uncover the connection between the speech perception and production modalities by training adult participants on a critical L2 vowel contrast in one modality and testing for transfer of learning effects in the other modality. The results from all three experiments taken together paint a detailed picture of the interaction and entanglement of the modalities during the process of learning second language sounds. In this chapter, I will first summarize the results of the three experiments and their component research questions. Next I will explain what the results say in terms of the underlying processes that occur during L2 speech perception and production and how the findings both align and misalign with current theories of L2 speech learning. Then, I will explore the limitations of this dissertation, and finally, I will close the chapter with implications for L2 methodology and pedagogy.

7.1. Summary of the main results

In this section I present a summary of the findings across the three experiments. For ease of comparison, Table 21 compiles all Cohen’s $d$ effect sizes and $p$ values from paired samples $t$-tests on the perception and production tasks from pre- and post-tests for the three trained groups of Experiments 1 and 2.

In Experiment 1, fifteen participants underwent a two-hour perception-only training on the English vowels /i/ and /ɪ/ and were tested on both perception and production. The results indicated that after training, participants significantly improved their perception to a large and statistically significant degree ($d=1.52$), which reached the level of the Bilingual Acquired performance, but not the level of the NS English group. They did not, however, improve their production. Although an acoustic analysis of the trained participants’ productions showed that
Table 21

Within group (pre-to-post) Cohen's d effect sizes and p-values from paired samples t-tests for the three trained groups on the perception and production tasks

<table>
<thead>
<tr>
<th>Task</th>
<th>Group</th>
<th>n</th>
<th>d</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification, shVp</td>
<td>Perception-Only</td>
<td>15</td>
<td>1.52</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>Production-Only</td>
<td>14</td>
<td>0.98</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td>No Sound</td>
<td>15</td>
<td>1.08</td>
<td>.02</td>
</tr>
<tr>
<td>Identification, fVt</td>
<td>Perception-Only</td>
<td>15</td>
<td>0.93</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>Production-Only</td>
<td>14</td>
<td>0.67</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td>No Sound</td>
<td>15</td>
<td>1.76</td>
<td>.001</td>
</tr>
<tr>
<td>Identification, natural</td>
<td>Perception-Only</td>
<td>15</td>
<td>0.62</td>
<td>.07</td>
</tr>
<tr>
<td></td>
<td>Production-Only</td>
<td>14</td>
<td>0.75</td>
<td>.04</td>
</tr>
<tr>
<td></td>
<td>No Sound</td>
<td>15</td>
<td>0.36</td>
<td>.13</td>
</tr>
</tbody>
</table>
| Elicited Production           | Perception-Only   | 15  | 0.29 | --   
|                               | Production-Only   | 14  | 0.84 | .00   |
|                               | No Sound          | 15  | 0.36 | .09   |
| Passage Reading               | Perception-Only   | 15  | 0.11 | --   
|                               | Production-Only   | 14  | 0.30 | .03   |
|                               | No Sound          | 15  | 0.27 | .08   |
| Picture Description           | Perception-Only   | 15  | 0.16 | --   
|                               | Production-Only   | 14  | -0.24| .13   |
|                               | No Sound          | 15  | 0.18 | .24   |

*Indicates that the repeated measures ANOVA did not reveal significant interactions of Time by Group by Task (p=.65), therefore, no further statistical tests were conducted to avoid type I error.

the difficult to produce /ɪ/ vowel became somewhat more target-like on the most controlled Elicited Production task, participants still did not show convincing evidence of producing the two vowels as distinct phonetic categories at post-test. Thus, Experiment 1 supported that perception training is beneficial for the development of L2 perception, a finding shown consistently in previous research (with an average of $d=.97$ across eighteen study samples, in Sakai & Moorman, submitted). The benefits of perception training did not, however, transfer to production.
Experiment 2 focused on two variations of a production-only training and once again tested for effects in both the production and perception mode. The two trained groups, consisting of twenty-nine participants total, underwent the same production-only training for two hours over four sessions, but the Production-Only No Sound group \((n=15)\) did not hear the sound of their own voices, whereas the Production-Only group \((n=14)\) experienced the training in a more ecological way by practicing the target sounds and hearing the normal auditory feedback produced by their own voices. Both production-only training conditions are rare in the literature in that (a) the training was visually rather than aurally supported, and (b) participants never heard other-produced auditory tokens of the target sounds. Most previous studies of production training have not focused on visual cues and have included perception models for analysis or imitation. The latter element means that, in effect, most previous studies have implemented a mixed perception-and-production training. In the present study, the Production-Only group consistently showed improvements in the production mode after training, while the participants in the Production-Only, No Sound group, who were denied the normal auditory feedback loop, did not improve their productions of the two target vowels. Specifically, when the NS judges’ ratings were used for examination, the Production-Only group improved their productions to a medium-sized \((d=.84)\) and statistically robust degree on the Elicited Production task, while the Production-Only, No Sound group showed a small and statistically non-significant improvement on this same task \((d=.36)\). The magnitude of production improvement after production training shown in the Production-Only group is similar to what has been previously reported in the literature for mixed production-and-perception training with access to other-produced auditory models (an average of \(d=.83\) from 110 within-group samples, in Lee et al., 2015). For the three previous studies that attempted to restrict participants’ other-produced
auditory examples of the trained phonemes (Herd, 2013; Liakin et al., 2013; Warsi, 2001), the effect on production after training was slightly smaller ($d=0.67$). When spectral overlap scores in the present production data were inspected, both trained groups improved, but the Production-Only group improved more. When acoustic cues were measured, both trained groups also improved their average /i/ and /u/ productions, but the Production-Only group again showed more target-like gains than the Production-Only, No Sound group. On the perception tests, both of the two trained groups improved their perception on the Identification task to a medium and statistically significant degree ($d=0.98$ and $d=1.08$, respectively). The Production-Only, No Sound group improved their identifications of both sets of synthetic stimuli to reach the level of the Bilingual Acquired performance but not the NS English group, and the Production-Only group improved their identifications on the synthetic fVt stimuli to reach the level of performance of the Bilingual Acquired group, but again, not the NS English group. Previous production training investigations that withheld other-produced auditory exemplars of the target phonemes and tested for effects in perception found a smaller effect of $d=0.51$ (Herd, 2013; Liakin et al., 2013; Warsi, 2001).

The comparison of the two different production-based training regimens in the present dissertation enabled the discovery of three important insights into the effects of access or no access to the auditory feedback loop during training. First, denial of access to the normal hearing auditory feedback loop provided an advantage during the production-only training (as shown in superior gains by the Production-Only, No Sound group from Sessions 2 through 4 of training; see Table 15 in Chapter 7) that did not translate to an advantage on the production post-test measures. Second, and as a corollary of the first finding, access to the auditory feedback loop during training provided an advantage on the production post-test, which was
done for both groups under normal conditions of access to one’s own auditory feedback. Finally, while the presence or absence of the auditory feedback loop seemed to have made a difference in the same-modality learning that was achieved, it did not influence the crossover effects from production into the perception mode at post-test.

Finally, Experiment 3 compared the pre-test performance on the English vowels /i/ and /ɪ/ of two baseline groups: a Bilingual Acquired group of L2 participants who were selected to serve as a baseline because they scored 80% or higher on both perception and production prescreen tests, and a NS group of participants who were selected to serve as a baseline because their L1 was English. The bilingual group comprised native speakers of Spanish who began learning English at an average age of six and had lived in an English speaking country at the time of testing for an average of eight and a half years. The native speakers of English had functionally monolingual childhoods and grew up in many parts of the United States. They professed to not have high proficiency in a second language nor did they extensively use an L2 in their daily lives. The two groups were found to perform similarly on the Identification task with naturally recorded tokens, on the Discrimination task with synthetic shVp tokens, and on the Elicited Production task, which elicited careful monitored speech. Differences emerged in perception on the Identification task with synthetic tokens and in production on the two tasks that elicited less monitored and more free speech, the Passage Reading and Picture Description tasks. In these cases, some participants in the Bilingual Acquired group showed they were vulnerable to a decline in performance, resulting in larger variation among the participants in this group. However, decline patterns were intermittent across tasks and participants, in that the bilingual participants whose performance declined on one task may have scored in the range of NS English scores on all of the other tasks. Moreover, both groups produced clearly distinct /i/
and /i/ vowels, although the Bilingual Acquired group’s productions were closer on the spectral continuum than the NS English group’s vowels on all three tasks. Thus, one can conclude the two baseline groups had target-like phonetic /i/ and /ɪ/ representations and deployed them in perception and production tasks, but the representations seem to have been more fragile and more vulnerable to stressors among the Bilingual Acquired participants.

**7.2. Speech training leads to same-modality gains**

In this dissertation, I employed a common and ubiquitous research design: Participants underwent training with the intention of improving the trained skill, in this case speech perception or production. In the end, both of the perception and production training of a difficult phonetic contrast in Experiments 1 and 2 were effective in supporting same-modality gains.

The perception-only training in Experiment 1 was short, just two hours, and utilized an identification task with feedback. The effect size on participants pre-to-post Identification scores was $d=1.52$. This means that the perception-only training led to a large improvement on the perception of the two English vowels. The substantive features of this training program and the subsequent findings are comparable to many previous studies. The meta-analysis by Sakai and Moorman (submitted) gathered data on eight strictly controlled perception training studies that focused on second language vowels (i.e., Counselman, 2010; Gomez Lacabex & Garcia Lecumberri, 2010; Lambacher et al., 2005; Lengeris, 2009; Nobre-Oliveira, 2007; Soler-Urzua, 2011; Thomson, 2007; and Wang, 2002). Of these articles and dissertations that reported the results in detail, the average within group effect size of perception-only training on the perception mode ranged from -.10 to 2.84, and the average effect size was 1.00. The perception-only training programs ranged from one to five hours in duration and included a wide variety of tasks, including identification tasks, discrimination tasks, rule-deducing activities, homework
assignments, and text-to-speech activities. Because there have been numerous studies that focus on L2 vowel perception-only training, the results from the present study were somewhat expected. The data from Experiment 1 adds additional information to this growing body of literature.

The production-only training in Experiment 2 yielded a medium effect on a post-test for the Production-Only group \(d=.84\) and a small, not statistically significant effect for the Production-Only, No Sound group \(d=.36\). Because the production-only training I employed was unique in that participants in both groups never heard other-produced auditory tokens of the target sounds, and participants in the No Sound group additionally never heard self-produced tokens of the target sounds, training was put to a more stringent test than in most previous studies. Additionally, there are few other studies to which these results can be compared. However, there are three production training studies outlined in detail in the literature review of this dissertation that did attempt to train participants in L2 segments without the use of auditory examples of the target phonemes (i.e., Herd et al., 2013; Liakin, Cardoso, & Liakina, 2013; and Warsi, 2001). Because all of these publications included detailed data reporting, I was able to calculate within-group, pre-to-post effect size scores for each of the trained groups in order to compare the findings to the present study. Liakin et al. (2013) found that participants significantly improved their productions of the French /y/ vowel after undergoing training through a phone application with an effect size of \(d=.74\). Warsi (2001) trained two groups on the English /ɔ/ - /ɹ/ distinction. The first received visual depictions of the vocal tract and articulators along with explanations of the articulation of the target phonemes. In a classroom with the other participants, learners practiced the sounds and received feedback from the instructor/researcher. It should be noted that although the instructor/researcher never provided
the participants in this study with examples to mimic, the participants did hear other-produced tokens of the target sounds produced by their classmates. The effect of this type of production training was $d=.82$. A second group of learners underwent the same training except this group did not receive the vocal tract illustration. The effect of this version of training was $d=-.04$. Finally, Herd et al. (2013) reported that participants learning L2 Spanish obstruents through waveform displays improved their productions with an effect size of $d=1.17$. To add perspective to the present findings against a larger pool of pronunciation focused investigations, I consulted the meta-analysis conducted by Lee et al. (2015). The authors collected data from 110 pre-to-post samples, and fifty-three of them targeted second language vowels. The average within-group effect size of the studies targeting vowels was $d=.91$. Because the pronunciation instruction meta-analysis was not concerned with strict separation of the modalities, it can be reasonably assumed that all of the studies offered access to self- and/or other-produced aural input throughout the training. The results from the meta-analysis and the aforementioned three studies taken together indicate that production training is often successful with second language learner samples. Experiment 2 in this dissertation augments this generalization by showing that even when access to good exemplars of the target sounds are withheld in order to make the training more purely production-based, production training improves production. It is notable, moreover, that the Production-Only group improved their productions very similarly to the average improvements reported in Lee et al.’s meta-analysis ($d=.84$ and $d=.91$, respectively) even though the participants in Experiment 2 never heard auditory examples of the target phonemes other than their own. Thus, it does not seem to be the case that the widely attested benefits of production training on production need to exclusively come from the opportunity
during production to perceive and process good exemplars of the target sounds that are typically embedded in many production training regimens.

The Production-Only condition aligns more closely to the natural conditions implemented in traditional pronunciation instruction studies and even closer to the training conditions with no other-produced exemplars by Herd et al. (2013) and Liakín et al. (2013). Therefore it is not entirely surprising that, of the two production-only training conditions in the present study, this condition was most successful in improving participants’ production. On the other hand, the No Sound training is unique and the first of its kind; there were no precedent studies to create a hypothesis of successful or unsuccessful outcomes. Thus, in uncharted waters, I waited for the results, and it appeared that participants were not able to improve their productions to a statistically significant degree, namely, in the ear of the native speaker judges who rated their production performances across the three production tasks. Participants in the No Sound group trained in an environment that was wholly different than their natural everyday lives, that is, without the sound of their own voices. In hindsight, it is perhaps less surprising that they were unable to show improvements when one considers that once they were post-tested, they again had access to their own voices. In other words, it may be that the discontinuity between training –where they did show improvement from session 2 to session 4– and testing conditions reduced the observed transfer of learning.

I suspect there are two additional factors that negatively influenced the pre-to-post performance of the two production-only training groups. First, as I mentioned in the Results Chapter, many participants lost a substantial portion of the first training session due to preparatory procedures for the Vowel Shapes program. In losing some or all of an already short training period, participants may not have achieved the full impact of what a true two hours
could have effected. Secondly, I used the Elicited Production task as the primary measure to reveal production gains at post-test. However, this task included a variety of minimal pairs on which the participants were not trained. Furthermore, they were not trained how to apply any newfound knowledge of /i/ and /ɪ/ production to new words. Had I used an Elicited Production task with only sheep and ship, the data might have revealed larger learning gains for both the Production-Only and the Production-Only, No Sound groups. This methodological feature will be explored further in Section 7.8 of this chapter.

Overall, the perception-only training showed improvements in the perception modality concerning a difficult L2 contrast, and one condition of the production-only training—the more ecologically valid one—showed improvements in the production modality of the same L2 sounds. The inclusion of the Production-Only, No Sound condition in the dissertation enabled two insights: (1) perhaps the auditory feedback loop is crucial in speech production learning, and (2) it is difficult to overcome the discontinuity of the absence of the auditory feedback loop in training and the presence of it in post-testing. It remains perhaps striking that the denial of access to the auditory feedback loop for the group that underwent the No Sound condition had a perceivable disruptive effect for the same-modality learning evidenced by post-test scores, whereas it was not detrimental to transfer of learning to the perception mode. In order to understand the same-modality learning, there are two parts of the procedure that are relevant: environment during the training and the environment of the post-test. During training, participants in the Production-Only group were able to create firm auditory and articulatory representations of the two vowels that they could call upon for later replication. This group of participants had access to two types of feedback (i.e., auditory and visual), which confirmed they were producing the correct vowel sound. In the testing condition, they still had access to
the auditory feedback loop, which could confirm they were producing the correct and intended vowel sound. On the other hand, the No Sound group only had access to one form of feedback during training (i.e., visual), and at post-testing, they no longer had access to this feedback to confirm the accuracy of the sound they were producing.

The participants in the No Sound group performed better on the difficult-to-produce /ɪ/ vowel during training when they had no access to the feedback loop. It seems that not hearing their own voices somehow aided them during training. In fact, participants in the No Sound group reported more often than the Production-Only group that they could feel a difference in the articulation for each of the vowels. They used lay terms to describe what they felt happening to their tongues and throats to make the vowel sounds (e.g., “I can feel the sound in the back of my throat.”). But then, the task at testing was very different than the training condition. Participants no longer had access to the visual feedback to reach the target and the category was less stable, and they were unable to sustain the successful productions during training.

7.3. Speech training can support cross-modal benefits

The driving motivation of this dissertation was the question of a connection between the perception and production modalities. In specific terms, I asked if one mode could be modified by training in the other.

The evidence suggested that overall perception-only training did not have a measurable effect on the production mode according to the NS judges’ identifications of the recordings on the three production tasks (cf. Experiment 1), but some small change appeared to be evidenced through two analyses. In other words, there is evidence of some change occurring in the production mode after perception, but the change was not substantial or apparent to listeners.
Although no transfer of learning from the perception mode to the production mode must be concluded for the findings in this dissertation, there is good reason to believe that learning does generally transfer in this direction. The meta-analysis by Sakai and Moorman (submitted) showed that the average effect of strictly controlled perception training on the production mode for twenty-four unique samples of participants was $d=0.54$ and reliably different from zero; the average effect size across eleven unique studies that targeted vowels was $d=0.46$, also statistically reliable. At the theoretical level, too, there are good reasons to believe that perception training should under optimal conditions support cross-modal benefits for production. Specifically, there is a large body of neurolinguistic literature that presents evidence that the motor areas of the brain are activated during auditory processing in difficult to perceive contexts. Some examples of challenging perceptual situations include listening in noise, listening to foreign accented speech (Callan, Callan, & Jones, 2014), and listening to non-native speech sounds (Callan, Tajima, Callan, Kubo, Masaki, & Akahane-Yamada, 2003; Wilson & Iacoboni, 2006). In ideal listening conditions, motor contributions may not be needed to perform regular discriminatory or identification tasks. However in the case of listening to non-native phonemes, the activation of the motor areas of the brain may aid listeners to perceive sounds better, especially in cases where they may be expected to produce the sounds later. This type of dual activation may lead to productive learning of L2 sounds after perception-only training.

I attribute the weaker results I found in this dissertation to three factors: length of training, target phonemes, and post-test measures. First, the average length of training in the meta-analysis was nearly five hours over the course of eight training sessions. The most famous perception training study that found effects in production is Bradlow et al. (1997), in which participants completed forty-five training sessions over a number of weeks. In the perception-
only training of the present study, I utilized a training regimen that was only two hours long, spread over four training sessions. It is possible that second language phonetic development is a process that takes more time than what I provided participants, especially when concerned with the transfer of skills to a second modality. Two hours of practice may have been just long enough for change to begin to occur but not long enough so that it could take full effect in the second modality. Secondly, there is evidence that manner of articulation may play a role in how easily perception learning transfers to production. In the meta-analysis by Sakai and Moorman, obstruents benefitted from the largest effect in production after perception training ($d=1.14$), followed by vowels ($d=.46$) and sonorants ($d=.39$). It is possible that the articulatory movements needed to make obstruent sounds are visually or gesturally salient in the auditory cue, thus allowing articulations to be learned through audition. In the same vein, the small articulatory movements necessary to make vowels and sonorants may be more difficult to “hear”. In this study, the articulatory difference of the two target vowels involves a very small change in the tongue root and body. It may be that these vowels in particular are not conducive to transfer from perception to production. Finally, transfer from the perception to the production mode may not have occurred because of a methodological feature of the research design. The production test that was most like the training was the Elicited Production task. However, this task may have been cognitively demanding since it required participants to produce various lexical items on which they did not train. Thus, the test required participants to apply knowledge learned in the training sessions to new phonetic and lexical contexts.

Turning to the opposite direction for cross-modal effects of training, the dissertation offered resounding positive evidence that strictly controlled production-only training transferred to the perception mode. Production-only training led to medium, statistically robust gains in
perception on the Identification task with synthetic shVp stimuli for both the Production-Only ($d=.98$) and Production-Only, No Sound groups ($d=1.08$). Additionally, both groups showed significant improvement on synthetic fVt tokens (Production-Only, $d=.67$ and No Sound, $d=1.76$), and the Production-Only group showed significant, medium-sized improvements ($d=.75$) on the naturally recorded tokens (here only, the No Sound group showed no evidence of change, with a small-sized and statistically not significant improvement, $d=.36$).

Of the three previously mentioned production-only training studies that tested for gains in the perception mode, just two provide comparability to the Production-Only group of the present dissertation. Recall that the training in Warsi (2001) took place in a classroom, where participants could hear their classmates producing the target phonemes. Because Warsi (2001) did not control for participants’ access to other-produced aural tokens of the target phonemes, this study cannot be included in this discussion of transfer from the production to the perception mode. However, Liakin et al. (2013) and Herd et al. (2013) did control for exposure to other-produced aural exemplars. In these studies, the researchers found a small effect on the perception modality after production-only training ($d=.49$ and $d=.57$, respectively). The Production-Only group of this dissertation matched the conditions of Liakin et al. (2013) and Herd et al. (2013), but the gain in the perception mode was two times as strong. Additionally, the condition that withheld access to the self-produced aural exemplars of the target phonemes maintained the strength of this effect. Thus, with the evidence of this dissertation, it cannot be said that the gains found in the perceptual mode after production training were due to perceptual activation and training by participants’ own voices. The present study thus provides for the first time robust evidence from a strictly isolated production training, with absolutely no access to any auditory examples of the target phonemes during training, that shows transfer of learning
from production to perception. In the next section, I explore the theoretical import of this finding.

7.4. Understanding production to perception learning

The neurolinguistic literature provides evidence that supports the claim that the auditory areas of the brain are recruited in speech production. However, since the No Sound group in the present dissertation also benefitted from cross-modal learning, it is reasonable to conclude that the perceptual gains attested were not driven by self-produced input during production training. Thus, another mechanism must be at play.

One potential explanation lies in predictive speech encoding which is a result of adults having years of prior experience associating articulatory gestures with their consequent sounds. Hickok (2012) explains, “the motor system can generate predictions for the sensory consequences of one’s own speech actions” (p. 7). In other words, as the vocal tract prepares to make a speech sound, and the auditory areas of the brain react, creating a prediction of the acoustic outcome of those physical gestures. As a sound is then produced, the predicted auditory model is compared with the sound that has been actually produced. If the predicted and the actual sounds match, no further concern is given to the process. However, if the predicted and actual sounds mismatch, the articulators will adjust in a new attempt to create the original intended sound. This is a pattern of predictive speech that is used to detect and correct production errors. This process, called the internal forward model of speech motor control, is represented in a long-attested line of literature (e.g., Aliu, Houde, & Nagarajan, 2009; Golfinopoulos, Tourville, & Guenther, 2010; Hickok, Houde, & Rong, 2011).

This mechanism of detecting and correcting errors may be particularly important for an L2 learner who is concerned with mastering new sounds. A study conducted by Parker Jones,
Seghier, Duncan, Leff, Green, and Price (2013) scanned sixty-seven native and non-native speakers of British English in an fMRI machine. As participants named pictures and read words in English, the researchers measured the amount of activation in the auditory and motor areas of the brain. They found that the non-native speakers had heightened auditory activation for their own speech compared to the native speakers. However, as the non-native speaker participants’ verbal fluency increased, the heightened activation for self-produced feedback began to look more like native speakers’ lower levels of activation. This may be indicative that the auditory areas of the brain in a non-native speaker are more alert for the presence of congruency across predictive to actual speech.

Callan, Callan, and Jones (2014) explain that it is hypothesized that “the establishment of those second language phonetic categories (when the second language is acquired after childhood) involves greater reliance on general articulatory-to-auditory feedback control systems, which generate auditory predictions based on articulatory planning” (p. 3). As the participants in both conditions of the production-only training prepared their mouths to make the /i/ and /ɪ/ vowels during training, the intended and probable auditory output was activated each time, regardless of whether they heard the auditory feedback or not. It is possible that experiencing the activation of the expected auditory output with each trial, hundreds of times over during the course of the training, strengthened the auditory representation of the target phonemes. Although the participants in the No Sound group did not receive confirmation through the auditory feedback loop that the predicted and actual sound matched, they did receive a type of confirmation in visual feedback from the Vowel Shapes program. The internal forward model of speech may be able to explain how production training can lead to perception
gains. The participants in this study showed that production and articulatory training alone can tap into the speech perception system.

7.5. Do the findings support the Speech Learning Model?

As discussed in the literature review presented in Chapter 2, the Speech Learning Model (Flege, 1995, 2003) posits that an accurate mental representation of an L2 phoneme is created by the ability to perceive distinctive features of the L2 phone. The SLM suggests that the L2 phonetic category contains intricate and detailed auditory and articulatory information located in long-term memory. The SLM does not make any claims about the existence of these or other types of stored phonemic or phonetic information. However, my careful reading of phonology literature suggests to me the mental representations assumed in the SLM likely entail stored categorical information that includes features such as phoneme frequency, phonotactic constraints, and lexical associations. In the process of second language speech learning, the L2 phonetic category becomes more fine-tuned through the ability to perceive the phone accurately, which later encourages accurate production of the sound. Thus, the SLM predicts that learning occurs in one direction, from perception to production. This dissertation questioned the relationship between perception and production implied in the SLM.

Firstly, this dissertation focused on the SLM claim that phonetic learning occurs through one vehicle: the perception mode. It did so by asking if the production mode can also be a vehicle that fine-tunes the phonetic category. The results from the No Sound group of Experiment 2 show that production-only training can in fact be a vehicle that leads to a more accurate mental representation of an L2 phoneme. The SLM does not predict the production-mode to be a catalyst of speech learning, yet participants who received production-only training significantly improved their perception from pre-to-post-tests. Thus, this study provides the first
evidence that the production mode can also drive learning of a target L2 phoneme, contra the predictions of the SLM. This finding needs to be challenged by more L2 phonology researchers who design strictly controlled production training programs that deny access to auditory examples of the target phonemes in order to replicate and further refine this finding. In addition, the field needs more evidence from a broad range of language combinations and phonemic targets to gather data about the potential of production-driven learning.

Secondly, this dissertation tested the SLM claim that there is a necessary order of L2 phonetic development: accurate perception appears before accurate production. It did so in two ways, first by analyzing gain scores of both modalities after training, and then through an analysis of the participants who were disqualified from the study due to test scores that rose above the 80% threshold to participate in training. The testing data gathered from the three trained groups in Experiments 1 and 2 showed evidence that perception gains are easier and more likely to be achieved than gains in production. The production mode is the more demanding of the two modalities, as it requires active planning, muscular and gestural control, and accuracy of an acoustic output. On the other hand, the perception mode is more passive in nature and simply requires hearing and categorizing sounds according to a mental representation. If the production mode is more demanding than perception, it is possible that practice in the harder modality may buy learning power in the less demanding modality. In this sense, the data from this dissertation upheld the SLM prediction that perception gains occur before production gains.

In addition, the data collected in this dissertation also makes it possible to directly address another element in the predictions of the SLM, namely, that accurate production never appears before accurate perception. This can be done by turning to the sixteen participants who were
disqualified from continuing with the training portion of the study because they scored higher than 80% on the production prescreen measure, but scored below 80% on the perception measure (See the Method Chapter). Henceforth, they will be called the “Production-High” group. At first glance, the existence of this group seems to contradict this SLM hypothesis. Table 22 provides all of the perception data from the participants in this group. A close inspection of the perception scores reveals that the participants actually scored quite high on the Identification task with the naturally recorded stimuli. The average score on this type of stimuli in the Identification task is 88%. This group of participants followed the pattern previously discussed for the Bilingual Acquired group, as all of these sixteen participants showed vulnerability on the synthetic stimuli in the Identification task. Thus, the average of the Identification scores was not able to show that this group could perceptually discern the difference between the two vowels on words that are produced naturally. This situation could be a result of the Production-High participants being sensitive to (and perhaps relying on) the subtle differences in duration that the NS model produced during the recording of these stimuli, or they have learned and mapped lexical examples of these vowels, but have not yet analyzed and synthesized the acoustic cues of the target phonemes to truly understand them at a more explicit level when called upon to judge the subtle differences between the synthetic tokens. Thus, this group that appeared to contradict the SLM, actually does not contradict it all. The data from the Production-High group is not evidence for or against the hypothesis that accurate perception necessarily precedes accurate production. There still remains an absence of evidence to prove or disprove this hypothesis.

7.6. A look at advanced learners who are motivated to overcome pronunciation difficulties
Table 22

Mean scores on the Identification task with three types of stimuli for the participants in the Production-High group

<table>
<thead>
<tr>
<th>Participant ID</th>
<th>Identification Task</th>
<th>Naturally Recorded</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>shVp</td>
<td>fVt</td>
</tr>
<tr>
<td>206</td>
<td>60.83</td>
<td>67.50</td>
</tr>
<tr>
<td>209</td>
<td>63.33</td>
<td>83.33</td>
</tr>
<tr>
<td>213</td>
<td>59.17</td>
<td>49.17</td>
</tr>
<tr>
<td>215</td>
<td>60.00</td>
<td>60.83</td>
</tr>
<tr>
<td>222</td>
<td>71.67</td>
<td>72.50</td>
</tr>
<tr>
<td>227</td>
<td>64.17</td>
<td>67.50</td>
</tr>
<tr>
<td>230</td>
<td>68.33</td>
<td>65.83</td>
</tr>
<tr>
<td>257</td>
<td>70.14</td>
<td>67.50</td>
</tr>
<tr>
<td>266</td>
<td>60.00</td>
<td>60.83</td>
</tr>
<tr>
<td>269</td>
<td>59.17</td>
<td>65.00</td>
</tr>
<tr>
<td>276</td>
<td>74.17</td>
<td>74.17</td>
</tr>
<tr>
<td>278</td>
<td>65.83</td>
<td>65.00</td>
</tr>
<tr>
<td>283</td>
<td>61.17</td>
<td>60.00</td>
</tr>
<tr>
<td>288</td>
<td>69.72</td>
<td>56.67</td>
</tr>
<tr>
<td>303</td>
<td>65.83</td>
<td>56.67</td>
</tr>
<tr>
<td>608</td>
<td>60.00</td>
<td>58.33</td>
</tr>
</tbody>
</table>

Note. Averages for each type of stimuli are presented on the last line of the table.

Participants in this study were highly skilled users of English who demonstrably were capable of meeting academic and professional demands requiring the advanced use of English on a daily basis. Thus, this dissertation offered a window into advanced language learners’ second language phonetic acquisition. This window is rare because most L2 perception and production training research has focused on participants with beginner and intermediate levels of proficiency. For example, the pronunciation meta-analysis conducted by Lee et al. (2015) reported that only 19% of 124 unique samples targeted advanced L2 learners, and Sakai and Moorman (submitted) reported that only one study out of eighteen focused on advanced learners. Moreover, the participants were not only a cross-section of advanced proficiency learners, but a
special one, targeted via a key element of the recruitment process: Potential participants responded to a flyer with an advertisement that explicitly asked if they wanted to improve their English pronunciation. In other words, from the outset, the recruitment narrowed in on one particular type of advanced learner, a learner who was motivated to improve his or her pronunciation. Self-selection undoubtedly influences the sample of participants in a study, and understanding the make-up of the people who volunteered to participate in this dissertation’s experiments is an important lens through which the findings should be interpreted.

After potential subjects contacted me expressing an interest in participating in the study, I asked them to fill out an online background questionnaire, which was used for two purposes: first, to screen for eligibility (e.g., being between eighteen and forty-five years of age), and second to collect linguistic and demographic information. Of the 143 NSs of Spanish who filled out the questionnaire, 95% stated they had a desire to improve their English pronunciation. (The remaining 5% may have been motivated to do the study for the monetary award or because of a friendly relationship with the researcher.) When asked how they felt about their current pronunciation abilities, the respondents replied with a broad range of thoughts. The most common response echoed a sentiment of feeling “ok” about the current state of their pronunciation, but knowing there was room for improvement. A few people said they felt good about the other communicative modalities (e.g., reading and writing), but that speaking continued to be a problem and they still felt a need to “travel a long distance” before reaching the goal. There were a number of emotionally driven responses. One person emphatically said, “My pronunciation is terrible!!!” Many people stated they felt “embarrassed because people don’t understand” them and another person summed up her response in one simple word, “Insecure.” Some answers revealed learner beliefs about pronunciation and accents, asserting
that it is something that needs to be eliminated: “I have a Spanish accent and despite so many years learning and studying English I cannot get rid of it,” and “I am very interested in working hard to totally remove any kind of accent I have.” Another common response was the feeling of being frustrated with their pronunciation, but not knowing how to improve this skill. Others related moments of intelligibility problems; one person said, people often “misinterpret what I say due to my pronunciation problems,” and another said, “[I] can’t tell the difference between ‘Yale’ and ‘jail,’ so there was some confusion when I visited some months ago.” In a subsequent question, 71% of respondents stated their pronunciation caused communication problems when speaking to others in English. In sum, while the participants who joined this study had a variety of motivations for desiring to improve their pronunciation, a majority professed issues of intelligibility in English communication.

Learners with high proficiency are important to investigate, observe, and train because results can inform researchers and educators alike on what aspects of L2 pronunciation are achieved without intervention and which elements persist in causing difficulty by the time learners reach the highest levels of proficiency. Figure 31 offers a visual summary of the participants who qualified and were disqualified from participation in the training because of prescreen/pre-test scores in each of the modalities. In order to qualify for the training of Experiments 1 and 2, participants had to score below 80% on the Identification task and below 80% on the productions recorded in the Elicited Production task. Because of this dual prescreening requirement, approximately 39% of those tested in the prescreening (n=44 out of a total 114 potential participants) did not qualify to continue with training because they scored at or above 80% in production (n=16), in perception (n=12), or in both (n=16). When considering perception alone, 24% of the entire potential participant pool (n=28 out of 114) scored above
80% on the average of all three stimuli sets in the Identification task. An additional 18% of participants who did qualify to join the training experiment (n=20 out of 114) scored 80% or more on at least one of the stimuli sets of the Identification task, which is indicative of unstable mastery of identification of the two target vowels. This means that in terms of perceptual categorical acquisition alone, 42% of potential study participants (n=48 out of 114) showed evidence of having acquired the /i/-/ɪ/ distinction in the perception mode, albeit some more extensively and stably than others. In terms of production alone, the productive categorical distinction of the two vowels had been achieved by 28% of potential participants (n=32 out of 114). An additional 6% of those who qualified to participate in the training (n=7 out of 114) scored 80% or higher on the NS judges’ identifications on at least one of the three production tasks (i.e., Picture Description, Passage Reading, and Elicited Production). This means that 34% of potential study participants (n=39 out of 114) showed evidence of accurately producing distinct /i/ and /ɪ/ tokens at pre-test.

Before the outset of this experiment, I incorrectly anticipated that very few of the recruited participants would be disqualified from the training portion of the experiment because of high scores. Based on the representation of the oft studied English /i/-/ɪ/ contrast (e.g., Escudero, 2006; Flege, 2003; Garcia Bayonas, 2008; Kondaurova & Francis, 2008, 2010; Mompean-Gonzalez, 2001; Morrison, 2006, 2009; and García Pérez, 2005), I understood these vowels to be difficult for all NSs of Spanish who learn to speak English, regardless of their overall proficiency or the goodness of their overall pronunciation. To confirm that I was not disillusioned in my prior thinking, I conducted a search in Google Scholar with the following search terms: Spanish, English, vowel, and second language. This search revealed forty-one studies that investigated Spanish learners of English vowels. Twenty-two of these studies
Figure 31. Pie charts that represent the percentage of participants who had high scores in each of the modalities at the prescreen-pre-test stage. Chart A presents the data from 114 participants at prescreen on

- Qualified for training: scored <80% on both perception & production
- Disqualified from training: scored ≥80% on only perception
- Disqualified from training: scored ≥80% on both perception & production
- Disqualified from training: scored ≥80% on only production

Chart B presents the data from 67 participants at prescreen on

- Qualified for training: scored <80% on all Identification stimuli
- Qualified for training: but scored ≥80% on at least one stimuli set
- Disqualified from training: scored ≥80% on perception prescreen test

Chart C presents the data from 76 participants at prescreen on

- Qualified for training: scored <80% on production tasks
- Qualified for training: but scored ≥80% on NS judges' identifications on at least one production task
- Disqualified from training: scored ≥80% on production prescreen requirements
both perception and production; B presents participants’ perception performance; and C presents participants’ production performance.

focused on /i/-/ɪ/ alone or among a small set of vowels, sixteen studies focused on all vowels, two targeted the low vowels /æ/-/ɑ/-ʌ/, and one focused on schwa. The English /i/ and /ɪ/ vowels are indeed a popular target to study for the Spanish speaking population, but perhaps the literature disproportionately suggests that it is a highly problematic sound contrast for Spanish speakers. Although this vowel contrast did cause difficulty for about 60% of the advanced learners in this study, the difficulty may not be as pervasive and persistent as the representation in the literature suggests.

In the end, this is a vowel contrast that needs intervention for the learners who –for reasons that await future investigation– have not acquired it by the advanced proficiency level. However, in terms of the literature, L2 phonological researchers should broaden the scope of target phonemes to include other vowel contrasts that may be equally as difficult for this population but understudied (e.g., the low back vowels /æ/-/ɑ/-ʌ/). An even better approach to select segments for research and training would be the Functional Load principle, which ranks vowel and consonant contrasts that carry the most weight for intelligible speech (Brown, 1988; Catford, 1987; Munro & Derwing, 2006). Not only would these suggestions deepen researchers’ knowledge of L2 acquisition of more portions of the vowel space, but it could ultimately help learners achieve success with key vowels.

Whether or not advanced learners respond well to training and intervention is another important empirical question. The present findings suggest that they do. But given the self-selected, highly pronunciation-motivated sample, further evidence must be sought from the portion of advanced learners who may feel satisfied with their levels of pronunciation and experience occasional or even frequent problems with intelligibility.
An inherent assumption in L2 phonetic training literature is that the onus of intelligible speech is on the non-native speaker. Many of the participants in this study showed evidence of taking this ideological stance, which Subtirelu (2014) calls the *deficit ideology*. If a learner believes the responsibility of clear communication is his or hers alone, a communicative breakdown may induce negative feelings, which could cause the learner to avoid future contact in the target language, leading to a halt in development. A different school of thought recognizes the contribution of the listener in achieving successful interactions. Subtirelu (2014) believes this view is the *lingua franca ideology*, where speaker and listener take on 50% of the responsibility for clear communication. Few researchers have approached communication from this standpoint, but those that do recognize its importance have focused on the role of the native speaker listener (Derwing, Fraser, Kang, & Thomson, 2014) and have even targeted training for this population (Derwing, Rossiter, & Munro, 2002). From this point of view, the learner does not assume blame for communication breakdowns, thus negative experiences are not internalized, and the learner will continue to seek out new linguistic experiences. LeVelle and Levis (2014) further emphasize this point, saying that social interactions can either be a “doorway or a wall” to language practice and progress. In any case, it is likely that at the advanced proficiency level, self-perceptions of pronunciation and intelligibility are particularly important for L2 users, since it is at this level that their many opportunities and demands for English language use may have enabled them to form strong (negative or positive) attitudes towards their own pronunciation. Based on the responses I received in the background questionnaire at the start of this study, it is likely that many of the non-native speakers who sought out an opportunity to improve their pronunciation carry the heavy weight of clear communication on their shoulders. The next logical question is to wonder how the actions of
researchers and educators can have an impact on learner beliefs and how those beliefs can be shifted towards more appropriate goals.

7.7. What baselines do we want for speech training studies?

In this study, I employed the use of two baselines, one bilingual group and one native speaker group. This design decision was a response to ongoing discussions in the L2 pronunciation literature about the value of intelligibility versus native-like performance (Thomson & Derwing, 2015) and more general discussions in second language acquisition regarding the charge of a monolingual bias in the research practices of the field (Cook, 1992, 1997; Ortega, 2014). Thus, the dissertation probes a second socio-cultural ideological stance that researchers can take in L2 phonetic training literature: to consider who should comprise a baseline group against which learners are compared. Functionally monolingual English speakers do not have two competing phonological systems, and thus will never be like bilinguals. And vice versa, the bilinguals will never be like the functionally monolingual (Cook, 1992, 1997). Lindemann, Litzenberg, and Subtirelu (2014) argue that maintaining the native speaker as the standard of comparison not only creates an abstract ideal for learners, but also sets an unrealistic goal. Instead, researchers and educators should focus on the “pronunciation of successful L2 speakers [as] a realistic and satisfactory goal” (p. 219). According to this line of thinking, bilinguals such as the Bilingual Acquired group in the present dissertation should be the ideal baseline for comparison, as they are bilinguals who have achieved the learning other L2 participants are striving for, in this case, the categorical distinction of the /i/ and /ɪ/ vowels in the perception and production modalities.

Comparing the two baseline groups to each other in the present study, and then comparing them to the trained groups of Experiments 1 and 2, was a valuable exercise that
offered a few insights. When I compared the two baseline groups to each other, they were never found to be statistically significantly different in the pairwise comparison tests I ran in Experiments 1 and 2. But when they were directly compared in Experiment 3, some differences arose. The Bilingual Acquired group consistently showed categorical acquisition on both modalities and on every task. But as the cognitive demands increased in the production and perception tasks, the bilinguals’ scores became more variable, as some participants were fragile or susceptible to vulnerabilities. Production seemed to be more tenuous than perception. When using the Bilingual Acquired group as a standard for the perception and production trainings, the three experimental groups that received training improved to reach the bilingual baseline on the Identification task at post-test with some, but not all, stimuli sets: The Perception-Only group reached the bilingual baseline on synthetic shVp tokens, the Production-Only group on synthetic fVt tokens, and the Production-Only, No Sound group on both the synthetic shVp and fVt tokens. Although the groups that received training were able to reach the Bilingual Acquired performance on some of the perception tasks, the trained groups never reached the NS English group scores on any of the tasks for perception or production. Using the Bilingual Acquired group as a baseline offers evidence of what bilinguals who have yet to acquire the English /i/-/ɪ/ distinction could reasonably achieve looking ahead in their learning trajectory, providing a useful benchmark for ultimate attainment in the acquisition of L2 sounds.

At this point it is important to note the differences in the demographic data between the Spanish-English bilinguals. Although I did not purposely sample the Bilingual Acquired group by any demographic characteristics, an inspection of Table 5 in the Method Chapter, reveals two differences between groups that may be useful for other L2 phonology researchers when they design their studies. One, the individuals who were assigned into the Bilingual Acquired
group were much younger when they began learning English (M=6.33 years old) compared to the other L2 groups (M=10.54 to 13.07 years old). Additionally, they were more uniformly and consistently younger, as shown in the fact that the variation of English onset (AO) for the Bilingual Acquired group was smaller (SD=3.02) than the other groups (SD=5.10 to 9.01). And two, the participants in the Bilingual Acquired group also had lived in an English speaking country for a much longer amount of time (M=8.47 years) than the other L2 groups (M=1.44 to 2.73 years), although in this case, the variation of length of residence (LOR) for the Bilingual Acquired group was large (SD=8.78), revealing the heterogeneity of their experiences. In fact, at the two extremes of LOR, four of the Bilingual Acquired participants had lived in an English-speaking environment for two years or less, while three participants had lived in an English-speaking country for more than eighteen years.

Taken together, these data support two inferences. First, it seems that the Bilingual Acquired group was composed of adults who had begun learning English at an earlier age. This pattern aligns with previous investigations in L2 phonology that an earlier start tends to be related to higher chances of a better outcome for pronunciation (e.g., Abrahamsson & Hyltenstam, 2009; Flege, Yeni-Komşhian, & Liu, 1999; Piske, MacKay, & Flege, 2001). Second, LOR does not necessarily predict individuals’ ability to produce L2 vowels. This pattern has also been substantiated by previous investigations (e.g., Flege & Liu, 2001; Moyer, 2011). LOR simply reflects “residing” in an L2 environment, but does not give information about actual quantity or quality of L2 use. In sum, the demographic profile of the Bilingual Acquired group as revealed by Table 5 in the Method Chapter is consistent with L2 phonology and SLA literature concerning individual differences. It is suggestive of the prevalence of age
effects in L2 phonological development, the low predictive value of LOR, and the importance to measure quantity and quality of L2 use in future research.

It may also be useful to future researchers to note that the mean age of English onset plus one standard deviation was nine years of age for the Bilingual Acquired group. If future researchers wish to sample bilingual participants who have likely acquired L2 vowel categories, recruiting participants with an AO of less than nine years of age may be an effective strategy to locate the intended population. On the other hand, this finding only emerged in post-hoc observations, and it should be subject to future replication. In the end, my purpose was not to investigate age effects or any other sources of individual differences, but the data suggest these factors are at play and are worthy of consideration in future research.

Although the field of SLA at large has been shifting away from holding the native speaker as ideal for over twenty years (e.g., Cook, 1999; Ortega, 2014; Bley-Vroman, 1983), the sub-discipline of L2 pronunciation has made slower progress. Thomson and Derwing (2015) composed a narrative synthesis of the same pool of studies included in Lee et al.’s (2015) quantitative meta-analysis of pronunciation instruction. They found that 63% of the studies “implicitly aligned with the nativeness principle,” which assessed participants against a native speaker target. Only 24% of the studies focused on goals of intelligibility, and the final 13% oriented to a mixture of the two ideologies. The authors of the synthesis do not say if there is a trend over time towards one ideology or the other. However, assuming that the field of L2 pronunciation is indeed making a shift away from using the native speaker as the ideal after years of advocacy from many prominent scholars in the field (e.g., Derwing & Munro, 1997, 2005; Levis & Moyer, 2014), looking forward, is there a need to have a NS baseline group at all? The direct comparisons of native speakers and bilingual participants undertaken in
Experiment 3 was informative in scrutinizing the value of the two baselines. The findings allow me to conclude that in terms of training and target-likeness in a study like this one, the Bilingual Acquired group is sufficient, and an appropriate point of comparison for L2 learners.

7.8. **Methodological lessons learned**

Reflecting on the methodological aspects of this dissertation from the privileged position of the other side of the finish line, there are two innovations of the design that I would recommend to future L2 phonology researchers and three changes I would implement were I to do this type of study again. The first innovation is the use of noise-cancelling headphones and the other is the vowel extraction technique I employed. The three changes I would make to the methodology pertain to the choice and design of pre- and post-tests, adjustments to the recruitment strategies to effectively reach outside the over-researched college populations and the production-only training conditions to make better use of the Vowel Shapes program. I will discuss each issue in turn, and then I will close the section with remarks about the costs (time and monetary) of this research.

This dissertation was innovative because it fore-fronted a type of methodology that is not typically used in second language research. I wanted to isolate participants’ articulatory gestures in the production-only training, so that they would not hear any exemplars of the target phonemes, including their own voices. This allowed me to fully disassociate production training from perceptual auditory experiences. In order to do this, I asked participants to wear noise-cancelling headphones and listen to low-pass filtered white noise (i.e., pink or brown). In this way, I was able to eliminate the air- and bone-conducted sounds of participants’ own voices. A logical hesitation may be that participants would resist this aspect of training, but no participants complained about this element of the study and no one dropped out after having
started the production-only training. (In actuality, more participants complained about the computerized, robotic voice of synthetic shVp and fVt tokens.) In fact, one participant said he used white noise to help him relax before bed, and he often yawned during the training, saying he felt very calm. This methodological feature has the potential to be very useful for researchers who are interested in disrupting the normal auditory feedback loop that participants with healthy hearing always have access to, for example, when interested in investigating a component of speech production and processing separated from audition. I highly recommend the use of this methodological feature.

A second aspect of my research design that has potential benefits for this domain of study is the vowel extraction technique that I employed. I chose to extract only onsets and vowels because I did not want the NS judges who identified participants’ productions to know the vowel the participant intended to say. Some of the produced words were part of a minimal pair. For example, participants produced the word meat, but, from the NS judges’ perspective, the participant could have been intending to say meat or mitt. This left the NS judges with a 50% chance of knowing the participants’ intended production. However, on most occasions, the words extracted for NS identification were not part of a minimal pair (e.g., wig, fish, consider, graffiti, eagle, and key). In these instances, if the NS judges had access to the whole word, their decision may have been influenced by knowing the intended utterance. In extracting just the onset and vowel, I allowed the coarticulatory information from the onset to remain obvious to the NS judges, but left the remainder of the word unknown. Identifications of the NS English group’s (n=20) productions verified that extracting just the onsets and vowels produced reliable scores. The NS English group’s vowels were consistently identified from 90% to 100%, and I took this as a sign that this extraction technique was reliable, while ensuring that the NS judges
would not be influenced by intended utterances. In the future, I recommend L2 vowel researchers who use a dependent measure involving judges to extract just onset and vowel information for identification. Moreover, I recommend this as an area for future research, as it is an empirical question if judges are influenced by the knowledge of a participant’s intended utterance.

After having completed all parts of the dissertation research, I would make three changes to the methodology. The first would be in the tasks I utilized for pre- and post-tests. I believe the Passage Reading task was not necessary and did not add much valuable information. This is because the Elicited Production task elicited monitored speech and the Picture Description task elicited unmonitored speech, and as a middle ground task, the Passage Reading did not provide any new information that the other tasks did not already reveal. Also, I would make one crucial adjustment to the Elicited Production task and have two sets of stimuli: ten repetitions of *sheep* and ten of *ship*, and then the set of minimal pairs that already comprised the Elicited Production task. I will discuss this issue further in the Limitations sections of this chapter.

A second methodological change I would make if given the opportunity is to rethink the efficiency of my recruitment strategy. I intentionally set out to recruit a variety of adults who were not necessarily students on a college campus. Many studies sample college students out of convenience, and this severely limits the generalizability of results. My intention was to survey adults holding a variety of professional positions in the real world. I succeeded in this regard, but it cost me nearly twelve months to complete the whole recruitment and data collection process. It was stressful. Every new recruitment idea I had yielded one or two new participants. Instead, in a future study, I would again prioritize not using college students as a sample, rather,
I would find and develop a relationship with a community-based school, church, or non-profit education center, and recruit from classes at one central location. If twelve students in multiple classes were incentivized to participate, I would have been able to recruit and collect data all at one location. Instead, the route I took was costly and emotionally draining. In the end, the most successful recruitment moment was when I made a contact in a D.C. Spaniards Facebook group. After the moderator posted about my study, a dozen new participants emailed me in twenty-four hours. I am indebted to this contact for all of the help that she provided.

The third methodological change I would make concerns the comparability of the two training designs. The perception-only training included four full sessions. On the other hand, the production-only training required quite a bit of time to calibrate the Vowel Shapes program, in terms of teaching the participants how the program worked conceptually (e.g., in order to move their dot higher, lift the tongue higher towards the roof of the mouth) and discovering each participant’s target phoneme location in the program’s interface. Because of these necessary precursors to using the production training program, many participants lost some or all of the first day of training, resulting in only three true full days of training. For a training program that is already very short, to lose up to 25% of it is quite significant and also affects the comparability across the training programs in Experiment 1 and 2. The effects of training are also compromised, meaning the production-only post-test scores could be much lower than what a true four-day training session could induce. Thus, if I were to conduct this experiment again, I would add an extra portion of time either at the end of the pre-test session or at the beginning of training Session 1 to accommodate the necessary time to become acclimated and adjusted to the Vowel Shapes program.
Finally, a note on the costs of this research is in order. Although valuable, training studies are taxing in the costs incurred, in terms of both time and money. Researchers wanting to conduct speech training studies may benefit from my experience as they develop plans and budgets for their study. First, recruitment, scheduling, and data collection were very costly. As mentioned, recruiting outside of the traditional four-year college environments was time consuming and demanding of creative strategies. In the end, recruiting and data collection took exactly twelve months and cost approximately $2000. I met with 135 participants in one-on-one thirty-minute sessions: the Trained participants met with me six times, the Control participants twice, and the remaining participants once. This translates to 374 thirty-minute sessions. Second, the cost of data analyses was equally taxing. The three human raters (one of them being the researcher) spent approximately fifteen hours each in order to judge eighteen thousand tokens (9,000 tokens, rated twice) produced by the participants. I spent approximately 250 hours preparing the acoustic data to be rated, and another sixty hours extracting formant values with a Praat script, hand-checking and correcting the output, and calculating the spectral overlap scores. Last but not least, creating the testing and training tasks, which included synthesizing vowels, took me approximately six weeks. In sum, I realize now I stepped into my study blindly, without actually knowing the cost and stress of completing a training study of this type. Prospective awareness of the costs will be invaluable for researchers undertaking their dissertations or writing grant proposals for the funding of studies similar to mine.

7.9. Study limitations

I have briefly referred to two limitations at certain points in this Discussion Chapter, and I will reiterate and expand upon them in this section. The first limitation of this dissertation involves the Elicited Production task used in the pre- and post-tests. The second limitation
concerns participants’ extra-experimental exposure and the consequent effects on the interdisciplinary contributions of this dissertation.

In order to capture all evidence of learning after speech training, a primary test should maintain as many substantive qualities of the training task as possible. For example, in Experiment 1, the perception-only training program was an identification task of synthetic sheep-ship tokens with feedback. To test effects of this training, I utilized an Identification task with synthetic sheep-ship tokens, and the test did not have feedback. Thus the primary test after perception-only training matched in task, stimuli quality (i.e., synthetic, not naturally produced), and phonetic context. A methodological concern expressed in Sakai and Moorman (submitted) suggests that in order for training programs to capture all evidence of learning and to be more consistent across studies, dependent measures should first be as similar to the training task as possible, and only then should they expand to subsequent tests to offer levels of generalization in task, stimuli quality, phonetic context, or any other additional change, as these tests of generalization are important to explore the ecological worth of the training.

In the planning and development of my pre- and post-tests for Experiment 2, I did not include a task that was most similar to the training. A primary production test for Experiment 2 should have been an Elicited Production task, using pictures to elicit responses, and prompting participants to say only the words sheep and ship. I did match the test to the training in task and elicitation procedure, but instead of maintaining the phonetic context and lexemes, I generalized the test words to eight sets of minimal pairs. Participants were not taught the orthographic equivalent of the /i/ and /ɪ/ vowels during training, and in my training and testing design, I intentionally avoided giving participants access to written words and attempted to only use images throughout the study unless otherwise unavoidable (i.e., the familiarization activity
before the Elicited Production task). Unless participants explicitly thought about the orthographic patterns of the two trained words, and then attempted to apply them to new words, there was little chance of them knowing which vowel was associated with which new lexical item presented in the Elicited Production task. A few participants asked me the orthographic equivalent of the sounds during training, and one participant discovered the pattern on her own. While speaking to herself aloud, she said, “If the pattern holds true, then all words with just the letter ‘i’ should be the [ɪ] sound, and the double ‘e’ words should be [i].” Even having discovered the spelling pattern, at the point of taking the Elicited Production task, this participant and all of the others would have had to see a picture, spell the word in his or her head, think of the orthography-to-sound pattern, and then produce the correct vowel sound. This is a complicated cognitive undertaking, and clearly does not represent the best option for the primary level of testing after a training regimen. Having a test of generalization act as the primary dependent measure, I may have missed the opportunity of capturing all evidence of production learning that occurred after production-only training and after perception-only training.

The second limitation in this study involves learners’ intentional and unintentional exposure to the target vowels outside of the experiment setting. Since the training was not intended to be implicit or incidental, throughout the four sessions, I informally talked to participants about their thoughts about pronunciation, the training program, and if they felt any changes happening in and outside of the experiment. Many participants spoke freely during their sessions, and one type of story reappeared over and over again: Participants were talking with family, friends, and coworkers about the research they were involved in. One male participant in the Production-Only group of Experiment 2 related that he asked his two
American-born daughters when he got home after the first training session to say the two sounds for him over and over again. Another participant in the Production-Only group asked her American coworker to help her with the sounds. Two participants in the No Sound training group independently said they practiced saying the sounds on the way to the lab and at night before bed. Another two participants in the Production-Only training group said they went out to a bar with Spanish friends and playfully argued about the sounds being the same or different. Finally, and most shockingly, a participant from Experiment 2 suddenly started performing very well during training. I said to her that I was happy that she was doing so well, but also wondered what had caused the change. She said that she was really happy she finally got it too and told me that she had been determined to figure out the sounds so she went home the previous evening after training and looked up YouTube videos on the target vowels. These are just the anecdotes that participants told me. I cannot begin to imagine the number of instances in which participants created an environment outside of the training that actively influenced the learning of the two target phonemes. L2 phonological development is not a fast process; it necessarily lasts for days, weeks, and months, and surely, in this time period, participants can go into the world and interact with the target phones in either modality.

Because many participants interacted with the target sounds outside of the experiment, I cannot say that the modalities were truly isolated during the period of training. This is not necessarily a complication for understanding the development of second language phonemes, as the perception and production modes are inevitably intermingled in real language use. However, I question if research using L2 training paradigms can legitimately contribute internally valid knowledge to the theoretical question of how the two modalities of perception and production interact in the L2 development. The speech training conditions implemented in this dissertation
were intentional, as they are in most of the speech training literature. That is, they did not in any case preclude participants’ paying conscious attention to the vowel contrast, which was the target of the training. If second language phonologists can develop a training regimen that succeeds in creating completely incidental learning conditions or a one-shot training program that ensures high internal validity, it may be possible to preserve the isolation of the modalities. If so, the ensuing results may be able to contribute information to the interdisciplinary understanding of the connection between speech perception and production. Otherwise, researchers may need to abandon the use of training studies for this purpose.

7.10. Pedagogical implications

The findings from this dissertation indicate three points relevant to L2 pedagogy: (1) many advanced learners still have difficulty with the English vowels /i/ and /ɪ/, whereas many have succeeded in acquiring this categorical distinction, (2) advanced learners show evidence of learning after perception-only and production-only training, and (3) focusing on either the perception or the production mode can benefit the other modality.

As previously discussed, perception and production of the vowels /i/ and /ɪ/ proved to be difficult for approximately 60% of the advanced English learners in this study. These advanced learners were able to show improvements in the perception and production modes after only two short hours of training, which suggests that even for the highly proficient, training and intervention can be effective. Although the scope of this dissertation did not include testing for the ideal moment in language proficiency to train the target vowels, results are able to speak to the advanced proficiency level, and suggest positive findings for intervention to occur at this stage, if necessary. The variability observed in participant performances at pre-test was substantial: 11% scored high on perception, but not production; 14% scored high on production,
but not perception; 14% scored high on both modalities; and 61% scored low on both modalities. This variability strongly suggests that the most efficient type of instruction may be learner-centered and corrective; in other words, instruction should target the individual needs of each learner. Thus, teachers interested in supporting the development of their advanced students’ L2 pronunciation in the classroom would be well served if they approached the decision to devote instructional time to the /i/ and /ɪ/ after ascertaining their actual students’ needs through diagnostic means, rather than assuming this contrast is or is not difficult, for example, based on the argument that the distinction is included in most pronunciation textbooks or is unknown in the students’ L1.

The final pedagogical implication of this dissertation offers optimism that time spent on one modality is likely to benefit the other, even if to a small degree. The findings from all five research questions of Experiment 1 and 2 taken together suggest that there is a complex interaction of the modalities. Ultimately, learners and instructors may be hopeful to hear that instruction in one modality can do double duty, as listening instruction will help listening and speaking, and speaking will help speaking and listening.
Appendices

Appendix A. Recruitment flyer

Example of one recruitment flyer targeting native speakers of Spanish

Research Participants
WANTED

Are you a native SPANISH speaker?
Do you want to improve your English pronunciation?

Earn $10 an hour doing 1-4 hours of listening and speaking tasks over a 2-week time period. The study takes place in the Education building on the USF campus.

Contact Mari at ms2335@georgetown.edu for more information.
Appendix B. Stimuli words and images

The following is a list of words and their associated images, which will be used for the perception identification training and tests.

<table>
<thead>
<tr>
<th>least</th>
<th>list</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="least Image" /></td>
<td><img src="image2.png" alt="list Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>reach</th>
<th>rich</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3.png" alt="reach Image" /></td>
<td><img src="image4.png" alt="rich Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>seek</th>
<th>sick</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image5.png" alt="seek Image" /></td>
<td><img src="image6.png" alt="sick Image" /></td>
</tr>
<tr>
<td>green</td>
<td>grin</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td><img src="image1" alt="Green" /></td>
<td><img src="image2" alt="Grin" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>deep</th>
<th>dip</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3" alt="Deep" /></td>
<td><img src="image4" alt="Dip" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sleep</th>
<th>slip</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image5" alt="Sleep" /></td>
<td><img src="image6" alt="Slip" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>cheap</th>
<th>chip</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image7" alt="Cheap" /></td>
<td><img src="image8" alt="Chip" /></td>
</tr>
<tr>
<td>leap</td>
<td>lip</td>
</tr>
<tr>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td><img src="image1.png" alt="leap" /></td>
<td><img src="image2.png" alt="lip" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sheep</th>
<th>ship</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3.png" alt="sheep" /></td>
<td><img src="image4.png" alt="ship" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>feet</th>
<th>fit</th>
</tr>
</thead>
</table>
Appendix C. Illustrations used for a production task

Illustrations used in the two counter-balanced versions of the Picture Description task
Appendix D. Embedded words in a production task

Images of the words below were embedded in the illustrations used for the Picture Description task. There were two counter-balanced versions.

<table>
<thead>
<tr>
<th>Version A</th>
<th>Version B</th>
</tr>
</thead>
<tbody>
<tr>
<td>/i/</td>
<td>/i/</td>
</tr>
<tr>
<td>jeep</td>
<td>mitten</td>
</tr>
<tr>
<td>bead</td>
<td>slipper</td>
</tr>
<tr>
<td>eagle</td>
<td>kitten</td>
</tr>
<tr>
<td>peach</td>
<td>racket</td>
</tr>
<tr>
<td>graffiti</td>
<td>chick</td>
</tr>
<tr>
<td>taxi</td>
<td>nickel</td>
</tr>
<tr>
<td>peanuts</td>
<td>sticker</td>
</tr>
<tr>
<td>candy</td>
<td>whistle</td>
</tr>
<tr>
<td>leaf</td>
<td>fifteen</td>
</tr>
<tr>
<td>teeth</td>
<td>pig(gy bank)</td>
</tr>
<tr>
<td>tea</td>
<td>tennis</td>
</tr>
<tr>
<td>leak</td>
<td>cigarette</td>
</tr>
<tr>
<td>sleep(ing)</td>
<td>bridge</td>
</tr>
<tr>
<td>key</td>
<td>six</td>
</tr>
<tr>
<td>three</td>
<td>big</td>
</tr>
<tr>
<td>peacock</td>
<td>iguana</td>
</tr>
<tr>
<td>leash</td>
<td>zipper</td>
</tr>
<tr>
<td>beetle</td>
<td>lizard</td>
</tr>
<tr>
<td>leaves</td>
<td>pickle</td>
</tr>
<tr>
<td>puppy</td>
<td>wig</td>
</tr>
<tr>
<td>ski</td>
<td>pitcher (of juice)</td>
</tr>
<tr>
<td>bee</td>
<td>tic-tac-toe</td>
</tr>
<tr>
<td>fifty</td>
<td>fifty</td>
</tr>
<tr>
<td>cheese</td>
<td>jacket</td>
</tr>
<tr>
<td>meat</td>
<td>pin</td>
</tr>
<tr>
<td>priest</td>
<td>chicken</td>
</tr>
<tr>
<td>heat(er)</td>
<td>fish</td>
</tr>
<tr>
<td>tree</td>
<td>stick</td>
</tr>
<tr>
<td>peace (sign)</td>
<td>office</td>
</tr>
<tr>
<td>week(ly calendar)</td>
<td>little</td>
</tr>
</tbody>
</table>
Appendix E. Passages used in a production task

The following passages were used in the Passage Reading task in the pre- and post-test. Red-colored text indicates words that contain the /i/ vowel, and blue-colored text indicates words that contain the /ɪ/ vowel.

Version A (206 words)

As Elizabeth had no longer any interest of her own to pursue, she turned her attention almost entirely on her sister and Mr. Bingley; and the train of agreeable reflections which her observations gave birth to, made her perhaps almost as happy as Jane. She saw her in idea settled in that very house, in all the felicity, which a marriage of true affection could bestow; and she felt capable, under such circumstances, of endeavoring even to like Bingley's two sisters. Her mother's thoughts she plainly saw were bent the same way, and she determined not to venture near her, lest she might hear too much. When they sat down to supper, therefore, she considered it a most unlucky perverseness which placed them within one of each other; and deeply was she vexed to find that her mother was talking to that one person (Lady Mitchell) freely, openly, and of nothing else but her expectation that Jane would soon be married to Mr. Bingley. It was an animating subject, and Mrs. Bennet seemed incapable of fatigue while enumerating the advantages of the match. His being such a charming young man, and so rich, and living but three miles from them, were the first points of self-congratulation.

Version B (217 words)

It was such a comfort to think how fond the two sisters were of Jane, and to be certain that they must desire the connection as much as she could do. It was, moreover, such a promising thing for her younger daughters, as Jane's marrying so greatly must throw them in the
way of other rich men; and lastly, it was so pleasant at her time of life to be able to consign her single daughters to the care of their sister, that she might not be obliged to go into company more than she liked. It was necessary to make this circumstance a matter of pleasure, because on such occasions it is the etiquette; but no one was less likely than Mrs. Bennet to find comfort in staying home at any period of her life. She concluded with many good wishes that Lady Mitchell might soon be equally fortunate, though evidently and triumphantly believing there was no chance of it. In vain did Elizabeth endeavor to check the rapidity of her mother's words, or persuade her to describe her felicity in a less audible whisper; for, to her inexpressible vexation, she could perceive that the chief of it was overheard by Mr. Darcy, who sat opposite to them. Her mother only scolded her for being nonsensical.
Appendix F. Instructions for the production training program

Participants received the following text on a handout on the first day of production training.

Instructions for Vowel Shapes

This program is designed to help people improve their vowels. As you say a word, it will track your voice and put a dot on the screen to represent your vowel. Each space on the graph represents a vowel area.

When the program is running, you will see a BLUE dot, which represents your TARGET vowel. Your voice will be represented by a red dot.

Your goal is to move the RED dot by changing the pronunciation of your vowel, and to try to reach the blue dot. Your red dot will move as you make slight adjustments to your tongue and mouth.

Once you’ve come close to the target, your dot will change to GREEN! Try to hold the vowel at that exact spot.

Practice saying some vowels now.

Say, “AAAAAAAAAAAA”

Say, “OOOOOOOOO”

Notice your red dot moves on the screen to different vowel spots.

When you do the pronunciation training sessions, you will look at two screens. The screen on your right will have the vowel shapes program. The screen on your left will have your instructions. The instructions will prompt you to say a word and to stretch out the vowel for a number of seconds. For example, the screen may look like this:

The picture shows you what word you should say, and the number indicates how long you should stretch the vowel sound in seconds.
The numbers will count down to show you how long to say the vowel. Then you will see STOP. The screen will look like this:

Next, it will say, to try it again.

Finally, it will say to take a rest and stretch your mouth and tongue out. So open your mouth wide and stick out your tongue!

After about 10 tries, you will take a break. You can rest, stand up, or stretch for about a minute. The researcher will change the vowel target, and then when you are ready, she will start the program again.

Total, the training will take about 20 minutes. You will get a few minutes at the beginning and end of the training session to practice freely with the program.

Let the researcher know you have finished reading this packet, and let her know if you have any questions.
Appendix G. Articulatory information used in the production training

Participants received the following handout on the second day of production training.
Appendix H. Compiled scores for all groups

A compilation of all trained, control, and baseline group pre- and post-test scores on the Identification task (A) and the three production tasks (B).

A

<table>
<thead>
<tr>
<th>Group</th>
<th>shVp</th>
<th>fVt</th>
<th>Natural Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Post</td>
<td>Pre</td>
</tr>
<tr>
<td>Perception-Only</td>
<td>50.17 (11.82)</td>
<td>71.89 (16.41)</td>
<td>48.90 (16.53)</td>
</tr>
<tr>
<td>Production-Only</td>
<td>49.70 (17.12)</td>
<td>66.07 (16.37)</td>
<td>62.50 (12.99)</td>
</tr>
<tr>
<td>No Sound</td>
<td>49.39 (18.87)</td>
<td>68.45 (16.48)</td>
<td>45.33 (20.70)</td>
</tr>
<tr>
<td>Control</td>
<td>44.69 (13.87)</td>
<td>52.67 (18.91)</td>
<td>50.47 (21.22)</td>
</tr>
<tr>
<td>NS English</td>
<td>85.88 (5.98)</td>
<td>--</td>
<td>88.63 (5.32)</td>
</tr>
<tr>
<td>Bilingual</td>
<td>81.72 (6.17)</td>
<td>--</td>
<td>83.59 (8.74)</td>
</tr>
</tbody>
</table>

B

<table>
<thead>
<tr>
<th>Group</th>
<th>Elicited Production</th>
<th>Passage Reading</th>
<th>Picture Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre</td>
<td>Post</td>
<td>Pre</td>
</tr>
<tr>
<td>Perception-Only</td>
<td>56.29 (16.84)</td>
<td>61.54 (19.74)</td>
<td>66.21 (18.43)</td>
</tr>
<tr>
<td>Production-Only</td>
<td>48.22 (14.35)</td>
<td>63.96 (22.30)</td>
<td>53.06 (18.22)</td>
</tr>
<tr>
<td>No Sound</td>
<td>53.15 (14.98)</td>
<td>59.94 (22.02)</td>
<td>57.38 (18.20)</td>
</tr>
<tr>
<td>Control</td>
<td>54.52 (18.15)</td>
<td>53.80 (17.19)</td>
<td>57.94 (18.63)</td>
</tr>
<tr>
<td>NS English</td>
<td>98.80 (1.93)</td>
<td>--</td>
<td>95.50 (5.07)</td>
</tr>
<tr>
<td>Bilingual</td>
<td>93.80 (7.66)</td>
<td>--</td>
<td>85.94 (12.25)</td>
</tr>
</tbody>
</table>
References


193


