

THE ACQUISITION OF PHONETIC CATEGORIES:  
AN ARTIFICIAL LANGUAGE LEARNING STUDY

Emily Moeng

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill  
in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the  
Department of Linguistics.

Chapel Hill  
2018

Approved by:

Elliott Moreton

Misha Becker

Elika Bergelson

Jeff Mielke

Katya Pertsova

Jennifer L. Smith

© 2018  
Emily Moeng  
ALL RIGHTS RESERVED

## ABSTRACT

Emily Moeng: The Acquisition of Phonetic Categories<sup>1</sup>  
(Under the direction of Elliott Moreton)

Part of learning a language includes determining what variation is meaningful and what variation is not meaningful. This dissertation presents a series of artificial language learning experiments to provide a timeline of early phonological acquisition in naïve adult learners. The core contribution of this dissertation is to propose a domain-general, two-stage model of distributional learning consisting of a **Bias Stage** followed by a **Sensitivity Stage**. Additionally, this dissertation will explore the relation that distributional learning holds with three factors, **attention**, **environmental context**, and **lexical acquisition**. Chapter 3 presents a set of experiments to make the core argument that distributional learning occurs in two stages. It is argued that the underlying mechanism behind distributional learning is not to directly warp the learner’s perceptual space, contrary to models which have been proposed. Chapter 3 will also examine the role of attention and its relation to distributional learning. Chapter 4 presents an experiment which investigates the relationship between environmental context and distributional learning. Results of this experiment will be used to support a one-stage model of allophony acquisition. Chapter 5 presents a set of experiments which explore the disparity between distributional learning and lexical acquisition.

---

<sup>1</sup> Stimuli, data, and analysis script codes used in this dissertation are stored at:  
<https://github.com/emoeng/Moeng2018-dissertation>

For F & S ♥

## ACKNOWLEDGEMENTS

I would like to start with the confession that I more or less ended up at the University of North Carolina by chance. I came for personal reasons, rather than searching for a department that best fit my research interests. I very much came into graduate school with an undergraduate mindset: I would memorize facts, finish my homework, and do as much of the assigned reading as I needed to in order to get the grade that I wanted.<sup>2</sup>

Despite stumbling across UNC by chance, I truly do not think I could have found a more welcome home than North Carolina, a better advisor than Elliott Moreton, or a more supportive group of friends than those that I made in the UNC linguistics department.

I want to start by thanking Elliott for challenging me and improving every aspect of my research. He showed me what it means to think like a graduate student, and I am incredibly grateful for the never-ending well of patience he showed me as I (very slowly) came to this understanding. Whenever I was ready to dismiss a dataset as “messy” and “uninterpretable,” I could always rely on him to connect the dots to previous research. He showed me that, despite the messiness that can come with studying human behavior, every dataset from a well-planned experiment is a theoretical puzzle that can yield something interesting, if you only look at it in the right light. I believe this organized thought process is the same reason I would often go into a meeting with him, feel as though I was tossing a jumble of half-formed

---

<sup>2</sup> I should probably specify that this was at least *my* mindset during my undergraduate years, which may explain why the undergraduates I have had the pleasure of teaching in my time here have been far better students than I ever was.

This material is based upon work supported by the National Science Foundation Doctoral Dissertation Research Improvement Grant No. 1451792. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

ideas in the air, and still somehow come out with a systematic plan of action. I will forever be proud and grateful to be able to call him my advisor.

I would also like to thank Jennifer Smith for her guidance throughout my time in graduate school. Whenever I find myself at a wall, whether while planning a lesson, preparing for a talk, or writing a paper, I think of how Jen would approach the problem. I feel incredibly fortunate that I was able to learn from her as a teaching assistant, and my goal as a teacher these last years has always been to demonstrate some of the same clarity and enthusiasm she brings to the classroom every day.


I am grateful to the rest of my committee members, Misha Becker, Erika Bergelson, Jeff Mielke, and Katya Pertsova not only for their insight, but also for their patience with me. I am still shocked to find my committee filled with these talented and accomplished linguists, and even more shocked that every one of them always gave me any support I asked for. Not once did I find them unavailable, and I count myself lucky to have had such knowledgeable committee members that were willing to help me every step of the way.

Many thanks to the joint efforts of Rachel Hayes-Harb, LouAnn Gerken, and Jessica Maye for digging up and letting me use their stimuli, and to Masaki Noguchi, Grant McGuire, Neil Macmillan, and Lawrence DeCarlo for their kind and helpful responses to my emails.

I also want to thank the friends I made during graduate school, Megan Gotowski, Xue He, Crandall Hicks, Yuka Mura, Anissa Neal, Kayla Vix, Bonnie Wang, and Hang Zhang for their support, as well as the members of the UNC P-side Research Group for listening to the many iterations of my project. And, although I only spent a small percentage of my graduate school time in California, I am indebted to my Google friends as the end stretch was by far the most stressful. Thank you to Niki Acker, Kelsey Kraus, Valerie O'Brien, Justin Rill, Özge Sarigul, and Paul Willis for keeping me sane during this time, and for providing statistical and theoretical discussions to keep my mind in my research.

This dissertation would not be possible without my family. It is because of them and the love of education they instilled in me that I am where I am today. Thank you for always being there for me. Your constant love and support mean everything to me.

And last, I want to thank Jen Boehm, Amy Reynolds, and Kline Gilbert. My grad school family has literally fought over who would “get” me when I lost my keys and needed a place to sleep; and yet, they have also made it painfully clear that this was only the toughest of loves during our many late-night Catan parties. Whether by convincing me that pie is a fruit, or by assuring me I had not hit rock bottom after stress-eating my advisor’s secret candy stash, these three have been dependably amazing. I have always been able to count on you all to go out of your way to comfort me when I was feeling sad, boost my morale when I was feeling inadequate, or buy All The Ferreros when I had something to celebrate. I never doubted that I would have somebody to talk to, no matter how trivial the topic, and I owe that to you all (sorry, “y’all”). I have many examples that illustrate how extraordinary these three are, but I think the best way to sum up my relationship with them is with this text message:



I NEED HALP WHERE ARE YOU

## TABLE OF CONTENTS

LIST OF TABLES .....	xiii
LIST OF FIGURES .....	xv
CHAPTER 1: INTRODUCTION .....	1
Introduction.....	1
Synopsis of the Proposal.....	2
Relationship of Overall Proposal with Attention.....	4
Relationship of Overall Proposal with Environmental Context.....	5
Relationship of Overall Proposal with Lexical Acquisition .....	6
Outline of the Dissertation .....	7
CHAPTER 2: INTRODUCTION .....	9
Introduction.....	9
Distributional learning .....	9
Motivation.....	9
Experimental Support .....	16
Models of Distributional Learning.....	18
Main Concepts in Signal Detection Theory.....	19
The Relationship between Choice Theory and logistic regressions.....	26



Sensitivity in Logistic Regressions .....	26
Response Bias in Logistic Regressions.....	27
Sensitivity and Response Bias in Logistic Regressions.....	28
Adapting Choice Theory to Same-Different Experiments.....	28
Summary of Analysis to be Used.....	29
Variations in Distributional Learning Experiments .....	29
The Relationship Between Natural and Artificial Language Learning.....	32
Research Questions and Structure of Dissertation .....	34
Summary .....	34
 CHAPTER 3: RESPONSE BIAS AND SENSITIVITY IN DISTRIBUTIONAL LEARNING, AND THE ROLE OF ATTENTION .....	 36
Introduction.....	36
Background.....	37
Distributional learning and Maye and Gerken (2000) .....	38
Web-based Experiments .....	40
Response Bias vs. Sensitivity .....	42
Distributional Learning in Functional Phonology .....	48
Research Questions and Summary of All “A” Experiments.....	61
Summary of Experimental Designs .....	62
Summary of Results.....	63
Experiment A1 .....	64

Stimuli.....	65
Procedure .....	68
Analysis .....	73
Results.....	75
Discussion.....	77
Experiment A2.....	77
Procedure .....	78
Results.....	78
Discussion.....	80
Experiment A3.....	80
Stimuli.....	81
Procedure .....	85
Results.....	85
Discussion.....	87
Discussion of Experiments A1-A3: Bias and Sensitivity .....	88
Current Proposal: A Two-Stage Model .....	88
“Tone” Experiments.....	98
Methods .....	100
Results: Experiment A2-Tone.....	100
Results: Experiment A3-Tone.....	102
Discussion.....	103

Discussion: Attention in Distributional Learning .....	103
Discussion: Distributional Learning Online .....	106
Discussion: Filler Trials .....	107
Conclusion .....	108
CHAPTER 4: DISTRIBUTIONAL LEARNING AND ALLOPHONY: A ONE STAGE MODEL OF ALLOPHONY ACQUISITION .....	111
Introduction.....	111
Background.....	112
Learning Complementary Distribution .....	112
A One- or Two-Step Model of Allophony Acquisition .....	115
Research Question .....	116
Methodology.....	118
Stimuli.....	119
Procedure .....	122
Participants.....	126
Results.....	128
Phone Test.....	128
Rule Test.....	135
Discussion.....	139
Comparison with Experiment A3 .....	140
Earlier Exposure Time .....	142

Unexpected Results of Filler Stimuli .....	144
Conclusion .....	144
CHAPTER 5: A GAP IN BOTH BIAS AND SENSITIVITY IN APPLICATIONS TO WORD	
LEARNING .....	146
Introduction.....	146
Background.....	146
A Gap Between Phone Discrimination and Lexical acquisition.....	147
A Possible Role of Sleep .....	152
Research Question and Summary .....	153
Experiment C1 .....	155
Methodology.....	155
Stimuli.....	155
Procedure .....	157
Analysis .....	162
Results.....	164
Discussion: Experiment C1.....	170
Experiment C2 .....	171
Methodology.....	171
Analysis and Results.....	176
Discussion: Experiment C2.....	181
Discussion.....	181

Conclusion .....	182
CHAPTER 6: DISCUSSION AND CONCLUSION .....	183
Introduction.....	183
Summary of Findings.....	183
Synopsis of the Proposal.....	187
Discussion and Further Research.....	190
Behavior of Filler Stimuli .....	190
Lack of Significant Findings in “C” Experiments .....	191
How “Linguistic” is Distributional Learning? .....	192
Phonetic Distance .....	193
Weaknesses and Future Research .....	193
Summary and Future Study .....	195
APPENDIX.....	196
REFERENCES .....	208

## LIST OF TABLES

Table 1. Effect of increased bias towards a “different” response compared to effect of increased sensitivity on a learner’s responses.....	47
Table 2. Possible end state ranking for bimodally-trained learner (left) and monomodally-trained learner (right). .....	53
Table 3. Predicted percentage of “different” responses in a <i>same-different</i> task based on the rankings given in Table 2. ....	56
Table 4. Initial state of the learner. ....	57
Table 5. End state constraint rankings after monomodal training (left) and bimodal training (right). .....	59
Table 6. Predicted percentage of “different” responses in a <i>same-different</i> task based on the rankings given in Table 5. ....	60
Table 7. Summary of key differences in all “A” Experiments. Complete summary of is given in Table 16. ....	62
Table 8. Summary of procedure in Experiment A1. ....	68
Table 9. Variables used in regression analysis for “A” Experiments. ....	73
Table 10. Summary of fixed effects in the mixed logit model in Experiment A1. ....	76
Table 11. Summary of fixed effects in the mixed logit model in Experiment A2. ....	79
Table 12. Summary of fixed effects in the mixed logit model in Experiment A3. ....	86
Table 13. Summary of fixed effects in the mixed logit model in Experiment A2-Tone.....	101
Table 14. Summary of fixed effects in the mixed logit model in Experiment A3-Tone.....	102
Table 15. Questionnaire responses regarding participants’ attention. ....	104
Table 16. Summary of stimuli, procedures, and results for all “A” Experiments.....	109
Table 17. Summary of procedure in Experiment B. ....	122
Table 18. Number of participants included in analysis per condition.....	127
Table 19. Results of GLMM for the phone test. ....	129
Table 20. Summary of follow-up contrasts testing specific hypotheses for the Phone Test.....	130
Table 21. Results of hypothesis testing within the context of the overall model for main effect of Distribution (regardless of PairType) for the Phone Test. ....	133
Table 22. Results of GLMM for the rule test.....	136

Table 23. Results of Rule Test for old trials and new trials. ....	137
Table 24. Variables used in regression analysis for Experiment C1, testing for bias and sensitivity. ....	162
Table 25. Results of GLMM for Experiment C1. ....	165
Table 26. Summary of follow-up contrasts testing specific hypotheses. ....	167
Table 27. Results of GLMM for Experiment C1, testing for an effect of sleep. ....	169
Table 28. Summary of follow-up contrasts testing for an interaction between Distribution and Sleep, Experiment C1. ....	169
Table 29. Summary of follow-up contrasts testing for a main effect of Sleep, Experiment C1. ....	169
Table 30. Results of GLMM for Experiment C2. ....	176
Table 31. Summary of follow-up contrasts testing specific hypotheses. ....	178
Table 32. Results of GLMM for Experiment C2, testing for an effect of sleep. ....	180
Table 33. Summary of follow-up contrasts testing for an interaction between Distribution and Sleep, Experiment C1. ....	180
Table 34. Summary of follow-up contrasts testing for a main effect of Sleep, Experiment C2. ....	181
Table 35. Summary of significant results in Experiments A-C. Asterisks indicate unexpected findings. ....	184

## LIST OF FIGURES

Figure 1. Schematic of overall proposal. ....	2
Figure 2. Experience-based perceptual warping account of phonetic category acquisition. ....	4
Figure 3. Relationship of overall proposal with environmental context. ....	5
Figure 4. A Context Stage account of allophony acquisition.....	6
Figure 5. VOT for velar oral stops in English-speaking data. Figure adapted from Lisker and Abramson (1964).....	14
Figure 6. Illustration of familiarization frequency of onsets of critical stimuli for Bimodal and Monomodal groups during the training phase of Maye and Gerken (2000). ....	17
Figure 7. Internal response.....	21
Figure 8. Internal state for a signal to which the organism has great sensitivity to, where sensitivity is represented by $d'$ .....	21
Figure 9. Criterion response defines four probabilities: for RealWord trials (top) the location of a participant's criterion defines the participant's hit rate and miss rate; for NonceWord trials (bottom) the location of a participant's criterion defines the participant's false alarm and correct rejection rate.....	22
Figure 10. Participant responses can be divided into four categories: hits, false alarms, misses, and correct rejections. ....	23
Figure 11. Internal state of a participant with a high criterion (top figure) compared to internal state of a participant with a low criterion (bottom figure).....	24
Figure 12. Calculation of $d'$ . ....	25
Figure 13. Figure on the right shows increased sensitivity to Signal and No Signal stimuli compared to the figure on the left.....	26
Figure 14. Figure on the right shows increased bias towards “yes” responses compared to the figure on the left. ....	27
Figure 15. Illustration of the familiarization frequency of onsets of critical stimuli for Bimodal and Monomodal groups during the training phase of Maye and Gerken (2000). ....	39
Figure 16. Illustration of the Sensitivity Hypothesis. ....	46
Figure 17. Illustration of the Response Bias Hypothesis. ....	46
Figure 18. Tableau illustrating the initial state in which *CATEG constraints are ranked high and PERCEIVE constraints are ranked low.....	49
Figure 19. An example of *CATEG demotion and PERCEIVE promotion. Figure from Boersma et al. (2003).....	49



Figure 20. Sample tableau for listener who has heard more [20 ms] tokens than [0 ms] tokens. Upon hearing a [0 ms] token, this listener will perceive it as being /20 ms/. .....	50
Figure 21. (left) Possible outputs for an input of [0] and [140] when all *WARP constraints including and below *WARP(60) are very low ranked. (right) Possible outputs for an input of [0] and [140] when all *WARP constraints including and below *WARP(80) are very low ranked. ....	52
Figure 22. Probability that an input of [0 ms] (top) and an input of [140 ms] (bottom) are categorized in each of the 8 possible prevoicing categories.....	55
Figure 23. First 300 ms of spectrograms of G1a, G4a, and G8a for stimuli created by the author (left), and for stimuli created by Jessica Maye and LouAnn Gerken (right). ....	67
Figure 24. Log odds of participants responding <i>d</i> for critical trials (left) and filler trials (right) in Experiment A1. ....	76
Figure 25. Log odds of participants responding <i>d</i> for critical trials (left) and filler trials (right) in Experiment A2. ....	79
Figure 26. First 500 ms of critical syllables S <sub>1a</sub> (top), S <sub>4a</sub> (middle), and S <sub>8a</sub> (bottom). ....	84
Figure 27. Log odds of participants responding <i>d</i> for critical trials (left) and filler trials (right) in Experiment A2. ....	86
Figure 28. Illustration of the Sensitivity Hypothesis (top) and Bias Hypothesis (bottom). ....	89
Figure 29. Looking times for infants in Maye et al. (2002). ....	92
Figure 30. Looking times for infants in Experiments 1 and 3 in Yoshida et al. (2010). ....	92
Figure 31. Illustration of proposed mechanism behind phonetic category acquisition. ....	96
Figure 32. Log odds of participants responding <i>d</i> for critical trials (left) and filler trials (right) in Experiment A2-Tone. ....	101
Figure 33. Log odds of participants responding <i>d</i> for critical trials (left) and filler trials (right) in Experiment A3-Tone. ....	102
Figure 34. Training distributions for Non-Complementary (left) and Complementary (right) conditions in Noguchi (2016). ....	114
Figure 35. Predicted results as amount of exposure increases. ....	117
Figure 36. First 500 ms of critical syllables S <sub>1a</sub> (top left), S <sub>3a</sub> (top right), S <sub>6a</sub> (bottom left), and S <sub>8a</sub> (bottom right). ....	121
Figure 37. Illustration of familiarization frequency for Bimodal-NonComp group (top-left), Monomodal group (top-right), and Bimodal-Comp group (bottom) during training. ....	124
Figure 38. The difference (in log-odds) of participants responding that critical Same Pairs are “different” and responding that critical Different Pairs are “different.” .....	131

Figure 39. The difference (in log-odds) of participants responding that filler SamePairs are “different” and responding that filler DiffPairs are “different.” .....	132
Figure 40. Main effect of Distribution for critical trials (top) and filler trials (bottom). .....	134
Figure 41. Results of Rule Test, old trials.....	137
Figure 42. Results of Rule Test, new trials. ....	138
Figure 43. Log-odds of participants responding “different” for critical trials at all three ExposureTimes in Experiment B and in Experiment A3. Error bars indicate standard error. ....	141
Figure 44. Comparison of the predicted change in sensitivity as amount of exposure increases (top; copy of Figure 35) and actual change in sensitivity as amount of exposure increases (bottom; copy of Figure 38). ....	143
Figure 45. Hypothetical scenario illustrating an early “hump” in sensitivity for Bimodal-Comp group which could be missed by the first ExposureTime tested. ....	144
Figure 46. Effect of filters (shown by the rectangle) on attention to each of the three planes (shown by different shades on the ball). Figure from Werker and Curtin (2005). ....	151
Figure 47. Comparison of Experiments A1/A3 (left) with Experiments C1/C2 (right). ....	154
Figure 48. Comparison of Experiments C1 and C2. ....	155
Figure 49. Initial 300 ms of spectrograms of G <sub>1</sub> a (left) and G <sub>8</sub> a (right) for critical stimuli. ....	156
Figure 50. Summary of procedure on each day. ....	157
Figure 51. Sound-meaning pairs presented during the word learning phase. ....	158
Figure 52. Summary of MatchedPairs and MismatchedPairs. ....	160
Figure 53. Summary of each phase in Experiment C1.....	161
Figure 54. Results from Experiment C1 split by PairType, critical trials.....	165
Figure 55. Results from Experiment C1 split by PairType, filler trials. ....	166
Figure 56. Bias results for Experiment C1.....	167
Figure 57. Sensitivity results for Experiment C1.....	168
Figure 58. First 500 ms of critical syllables S <sub>1</sub> a (left) and S <sub>8</sub> a (right). ....	172
Figure 59. Summary of procedure on each day. ....	173
Figure 60. Sound-meaning pairs presented during the word learning phase. ....	174
Figure 61. Summary of MatchedPairs and MismatchedPairs. ....	174

Figure 62. Summary of each phase in Experiment C2.....	175
Figure 63. Results from Experiment C2 split by PairType, critical trials.....	177
Figure 64. Results from Experiment C2 split by PairType, filler trials. ....	177
Figure 65. Bias results for Experiment C1.....	179
Figure 66. Sensitivity results for Experiment C1.....	179
Figure 67. Breakdown of results from Experiment C1, critical trials.....	187
Figure 68. Schematic of overall proposal. ....	188
Figure 69. Experience-based perceptual warping account of phonetic category acquisition. ....	189

## CHAPTER 1: INTRODUCTION

### 1. Introduction

Despite being exposed to great variation and little to no explicit instruction, infants acquire linguistic structures with remarkable ease. Heart rate studies show that this acquisition begins even while infants are still *in utero* (DeCasper and Fifer, 1980; DeCasper and Spence, 1986; DeCasper et al., 1994). Newborns show recognition of their native language's prosodic structure over other prosodic structures (Mehler et al., 1988) and a preference for their mother's voice over other female voices (Mehler et al., 1978; DeCasper and Fifer, 1980). Infants also show a preference for human speech over primate vocalizations or human speech played in reverse (Dehaene-Lambertz et al., 2002; Pena et al., 2003; Vouloumanos and Werker, 2004, 2007). By the time an infant becomes a year old, they exhibit language-specific discrimination of sounds which are contrastive in their language (Eilers et al., 1979; Kuhl et al., 1992; Eimas et al., 1971; Kuhl et al., 2006). And even though infants initially acquire new words slowly, having an estimated receptive vocabulary of about 60 words and a productive vocabulary of only about 14 words by the age of 12 months (Bergelson and Swingley, 2015; also see Caselli et al., 1995; Ferguson et al., 2015), at around 18 months of age many infants display what is known as a "vocabulary spurt," producing as many as 60 new words in a 2.5 week period (Goldfield and Reznick, 1990; although see Ganger and Brent (2004) who argue that most infants do *not* undergo a vocabulary spurt).

The goal of this dissertation is to provide a **timeline of segmental acquisition**, from the early stage of simple perception to the later stage of utilizing acoustic distinctions in a semantically meaningful way. Conclusions will be based on experimental evidence from artificial language learning tasks conducted on adult learners, and, when available, evidence that suggests a valid extension to infant learners will be discussed. It is hoped that this dissertation will provide a model of early segmental acquisition

which can be further tested, as well as set a foundation for future infant language learning studies. In particular, the timeline of acquisition of allophony presented in Chapter 4 would likely greatly benefit from further work on infants.

## 2. Synopsis of the Proposal

The overall contribution of this dissertation is to identify and detail two main, domain-general stages of early phonetic category acquisition: a Bias stage followed by a Sensitivity stage.

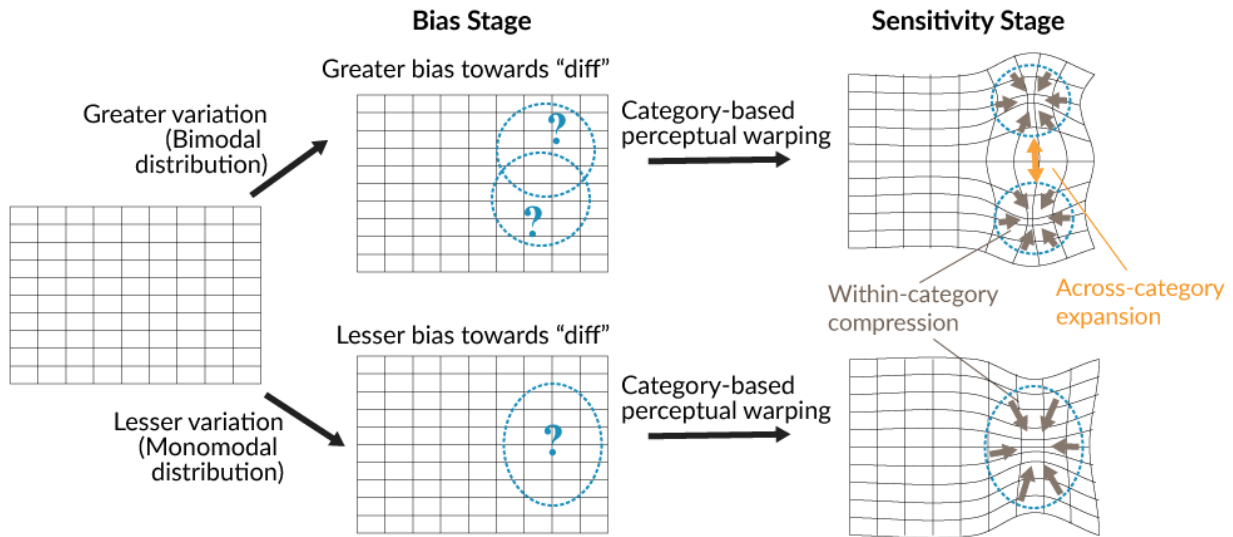


Figure 1. Schematic of overall proposal.

The initial stage of acquisition is represented by the leftmost grid in Figure 1. This grid represents the learner's perceptual space. Individual speech tokens can be mapped into this space, where the spatial distance between two given tokens represents perceptual distance. In reality, this initial grid is already warped by the learner's auditory system (Eimas et al., 1971; Kuhl and Miller, 1975, 1978; Kuhl, 1981; Aslin and Pisoni, 1980b) and, if there is one, existing L1 background (Aslin and Pisoni, 1980b; Cristia et al., 2011), but for simplicity will be shown as an evenly-distributed grid. This dissertation proposes that language learners who notice greater variation will come to expect more sound categories in the speech stream compared to learners who do not notice as much variation (Chapter 3). This in turn leads to a

change in participant **response bias**. This is indicated in the first stage, the Bias stage, in Figure 1. As indicated, the learner who notices more variation will hold some rough notion that multiple sound categories (two, in the upper succession of grids in Figure 1) exist within their perceptual space. At this early stage, the boundaries and the centers of these categories are unknown to the learner, as represented by the dotted lines. Likewise, the learner who does not notice as much variation holds the rough notion that only a single sound category exists (as shown in the lower succession of grids in Figure 1). Again though, the boundaries and center of this category are unknown to the learner. The different number of sound categories expected by each of these two learners results in a greater **bias** towards a “different” response in the bimodally-trained learners compared to the monomodally-trained learners. As learning progresses, learners form more solid hypotheses of the space each sound category occupies within their perceptual map, and experience category-based perceptual warping in a Sensitivity stage. During this stage, acoustic members deemed by the learner to belong to the same category are perceived as being more similar to one another in **within-category compression**, and members deemed by the learner to belong to different categories are perceived as being more different from one another in **across-category expansion**. These phenomena have been documented in a number of studies within psychology (e.g. Livingston et al. 1998; Goldstone and Hendrickson, 2010).

This two-stage proposal contrasts with a one-stage experience-based perceptual warping proposal illustrated in Figure 2.

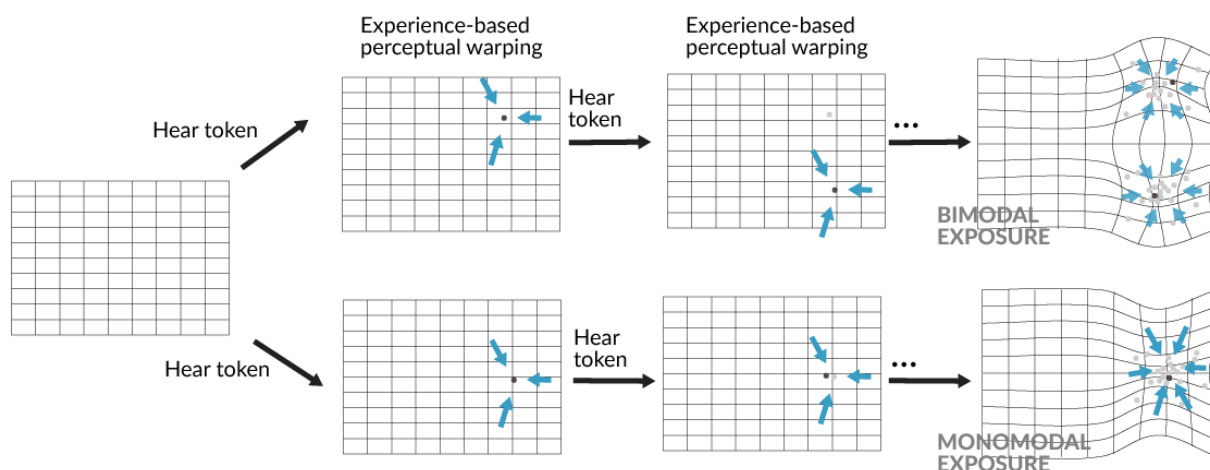


Figure 2. Experience-based perceptual warping account of phonetic category acquisition.

In this proposal, the experience of hearing some acoustic token warps the learner’s perceptual space (Guenther and Gjaja, 1996; Boersma et al., 2003). According to this class of models, the act of perceiving a token warps the perceptual space towards the token’s mapped location. If two clusters of tokens are experienced, as shown in the final stage of the topmost succession of grids in Figure 2, this results in two “centers of gravity” within the learner’s perceptual space. If only one cluster of tokens is experienced, as shown in the final stage of the bottom succession of grids, the learner will have one “center of gravity” in their perceptual space.

The main contribution of this dissertation is to argue for the model illustrated in Figure 1 for early segmental acquisition. Evidence supporting this proposal will be laid out in Chapter 3. Additionally, this dissertation aims to explore the relationship between this overall proposal and three elements: learner’s attention, environmental context, and lexical acquisition. Conclusions regarding each of these will be briefly summarized below.

## 2.1. RELATIONSHIP OF OVERALL PROPOSAL WITH ATTENTION

Two experiments presented in Chapter 3 provide evidence that learners’ attention plays some type of role in early phonetic category acquisition. Specifically, this chapter argues that attention plays a role in the *magnitude* of the arrows shown leading up to the Bias stage in Figure 1, and possibly also the magnitude

of the arrows leading up to the Sensitivity stage. In other words, this chapter argues that attention either aids or hinders phonetic category acquisition. Both of these suggestions are considered in this dissertation, but further research will be needed to determine the exact nature of the role that attention plays.

## 2.2. RELATIONSHIP OF OVERALL PROPOSAL WITH ENVIRONMENTAL CONTEXT

Chapter 4 explores the relationship of phonetic category acquisition with environmental context by mapping participants' early learning trajectories. Dillon, Dunbar, and Idsardi (2013) put forth a model of allophony acquisition that occurs in a single stage. This counters proposals such as Peperkamp et al. (2003), which models the acquisition of phonetic categories and the allophonic relationships between those categories as two separate stages (not to be confused with the Bias-Sensitivity two-stage model of phonetic category acquisition shown in Figure 1). This chapter presents evidence supporting a one-stage model of allophonic acquisition, suggesting that environmental context is taken into account by the learner from the very beginning of acquisition, and not in a second stage. This is schematized in Figure 3.

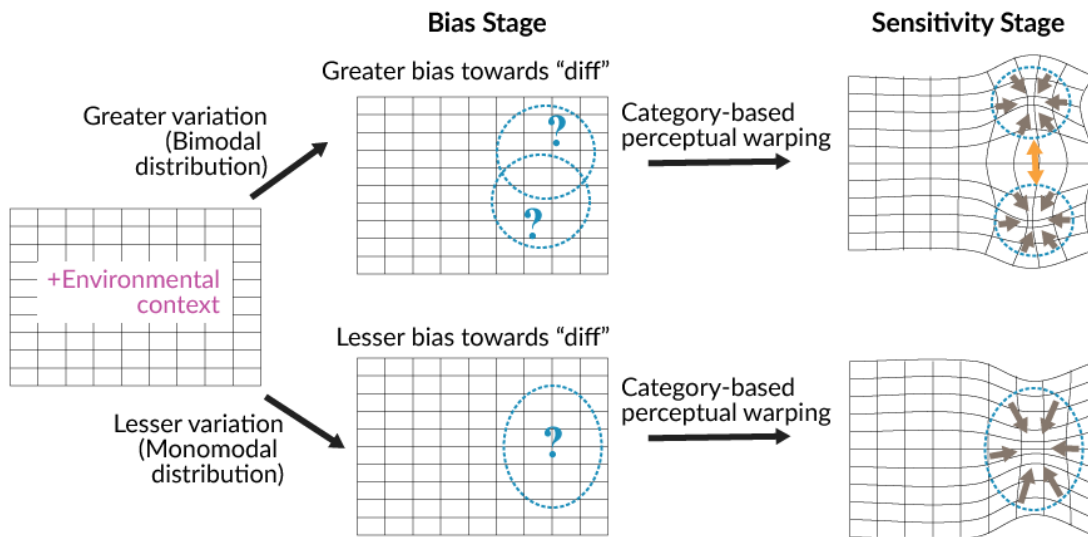


Figure 3. Relationship of overall proposal with environmental context.

This model contrasts with a model in which allophonic relationships are acquired in a separate stage of acquisition, a Context Stage, as shown in Figure 4. In this model, phonetic categories are acquired first (according to this dissertation, this is further broken down into a Bias Stage and a Sensitivity stage), and



allophonic relations are acquired in a following stage. In this hypothetical Context Stage, phonetic categories which occur in complementary environments are collapsed into a single phoneme, and phonetic categories which occur in contrastive environments remain as distinct categories. This dissertation finds no evidence supporting a Context Stage.

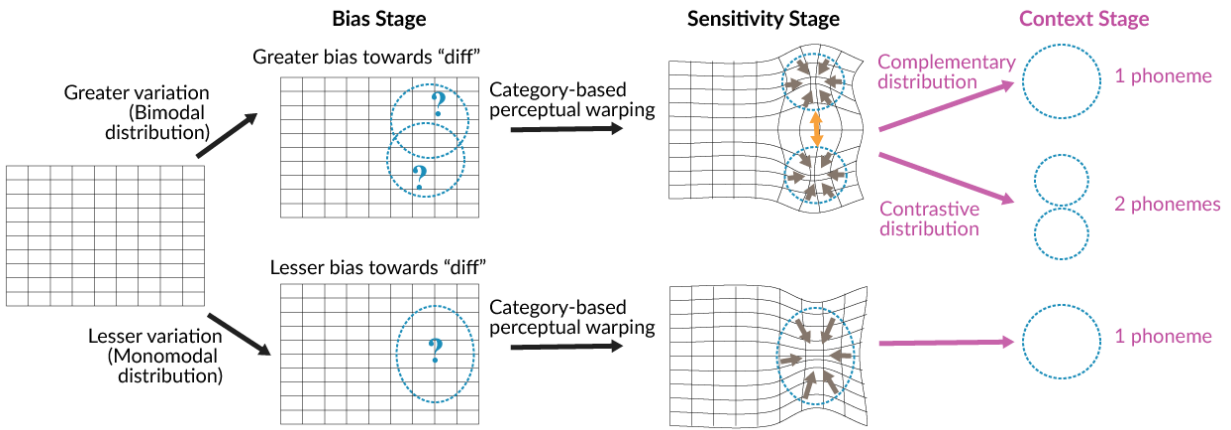


Figure 4. A Context Stage account of allophony acquisition.

### 2.3. RELATIONSHIP OF OVERALL PROPOSAL WITH LEXICAL ACQUISITION

Finally, this dissertation tests for whether any evidence can be found for the incorporation of response bias or sensitivity in a word-learning task. Previous research from L1 acquisition (Stager and Werker, 1997; Pater et al., 2004), L2 acquisition (Daidone and Darcy, 2014), and artificial language learning tasks (Hayes-Harb, 2007) all suggest that increased discrimination does not necessarily translate into use of some distinction in lexical acquisition (although see or Fennell and Waxman (2010); Rost and McMurray, 2009; 2010). Chapter 5 builds on a previous artificial language learning task (Hayes-Harb, 2007) to determine whether a period of sleep and more exposure to stimuli will result in an extension of either response bias changes or sensitivity changes to lexical acquisition. I find no evidence that changes in either response bias or sensitivity extends to a word-learning task, even over the course of three consecutive days of training. Findings from this chapter suggest the need to extend previous explanations of a gap in discrimination and word learning (e.g. Pater et al., 2004; Werker and Curtin, 2005).

### 3. Outline of the Dissertation

The remaining chapters present the specifics of the timeline described in the previous section. Chapter 2 provides the reader with background information and motivation regarding the current study. This includes a discussion of what is known as distributional learning, an oft-cited method used by learners in phonetic category acquisition. Previous models of distributional learning which assume a perceptual warping account of acquisition, as schematized in Figure 2, are also presented in this chapter. This is followed by a summary of basic concepts in Signal Detection Theory and the related concept of Choice Theory.

Chapter 3 introduces a distinction between **bias** and **sensitivity**, which will make up the two stages of the overall proposal schematized in Figure 1. Chapter 3 then presents three main experiments, Experiments A1-A3. These will be used to argue for the overall proposal given in Figure 1. Experiments A2 and A3 are followed by two “Tone” experiments, A2-Tone and A3-Tone, which suggest some role played by listeners’ attention during early phonetic category acquisition. In addition to presenting evidence for the overall proposal in Figure 1 and finding evidence that attention plays some role in early acquisition, Chapter 3 also concludes that distributional learning experiments can be replicated through web-based platforms, and not just in a typical laboratory setting.

Chapter 4 presents Experiment B, which explores the relationship between phonetic category acquisition and environmental context. Experiment B maps the learning trajectories of learners trained on one of three statistical distributions in order to determine whether evidence for a separate Context Stage, as shown in Figure 4, can be found. Learners are exposed to either 5 minutes of training, 10 minutes of training, or 15 minutes of training.

Chapter 5 presents the “C” Experiments. Two experiments, C1 and C2, explore whether the various components learned through distributional learning (changes in bias and sensitivity) extend to a word learning task. These experiments train learners over the course of three days in order to determine whether a period of sleep results in the incorporation of distributional learning in a word learning task.

Experiment C1 tests for evidence that a change in bias extends to word learning, while Experiment C2 tests for evidence that a change in sensitivity extends to word learning.

Chapter 6 highlights main findings and summarizes conclusions that were made in previous chapters. It ends with a discussion and suggestions for further research.

## **Chapter 2:**

### **Background Research and Motivation**

#### **1. Introduction**

The main topic under investigation in this dissertation is a statistical learning process utilized by language learners known as **distributional learning** (Maye and Gerken, 2000; Maye et al., 2002). This dissertation explores various characteristics of this process in order to propose a model of the underlying mechanism that drives distributional learning. This chapter will introduce distributional learning in Section 2, and detail the first experimental support for distributional learning, Maye and Gerken (2000) (Sections 2.1-2.2). This will be followed by descriptions of models of distributional learning in Section 2.3. Section 3 briefly detours away from distributional learning to describe the main concepts of Signal Detection Theory and the related Choice Theory (Luce, 1959), which this dissertation bases its analysis on. Following this, Section 4 returns to the question of phonetic category acquisition in a comparison of past studies of distributional learning. It will be highlighted that seemingly small variations in both methodology and analysis methods measure different aspects of distributional learning. This chapter ends with an acknowledgement of methodological problems in studying distributional learning in artificial language learning studies, while also justifying this dissertation's experiments in Section 5.

#### **2. Distributional learning**

##### 2.1. MOTIVATION

Infants show language-specific discrimination of vowels around 6-8 months (Polka and Werker, 1994; Kuhl et al., 1992; Trehub, 1976), and language-specific discrimination of consonants at 6-12 months (Eilers et al., 1979; Aslin and Pisoni, 1980a; Eilers et al., 1980; Iverson et al., 2003; Werker et al., 1981;

Werker and Tees, 1983, 1984; Werker and Lalonde, 1988; Best and McRoberts, 1989; Best et al., 1988; Trehub, 1976). Although adult Japanese speakers experience considerable difficulty distinguishing between [ɹ] and [l] (Iverson et al., 2003), 6 month-old Japanese infants can discriminate between these two sounds (Kuhl et al., 2006). Similarly, although adult English speakers experience difficulty distinguishing between [t] and [ʈ], 8 month old English-learning infants are still able to distinguish these sounds (Werker and Tees, 1984). These observations lead to the claim that infants are “citizens of the world” (Gervain and Mehler, 2010; Kuhl, 2004), having the ability to distinguish all contrasts which are linguistically-relevant. In this view, “acquisition” essentially equates to a loss of contrasts which are not linguistically relevant in the language being heard (Eimas, 1978; Morse, 1978, Werker et al., 1981; Gervain and Werker, 2008; Gervain and Mehler, 2010).

Several observations suggest a picture which is more complex than this simple “citizens of the world” view of language acquisition. First, some boundaries which fall between contrasting phones appear to stem from the auditory system, as evidenced in non-human studies and studies with very young infants. Chinchillas (Kuhl and Miller, 1975, 1978) and macaque monkeys (Kuhl and Padden, 1982), as well as 1-4 month olds (Eimas et al., 1971), show a greater ability to distinguish a pair of sounds which straddle a VOT boundary of 20-50 ms<sup>3</sup>, compared to an equally-spaced pair of sounds which do not straddle this same VOT boundary. This boundary corresponds to the location that many languages, including English, use as a boundary to contrast phoneme pairs such as /p/ and /b/. This suggests that at least some pairs of phonemes are separated by a natural acoustic boundary (similarly, see Kuhl and Padden, 1983).

Second, there are several exceptions to the seemingly “universal” discrimination which infants appear to exhibit from birth. For example, English makes a contrast between /d/ and /ð/ while French does not. However, both English and French 6-8 month olds show poor discriminatory ability between [d] and [ð] (Polka et al., 2001; Sundara et al., 2006). Narayan et al. (2010) find an effect of acoustic salience, with

---

<sup>3</sup> The actual boundary value depends on place of articulation.

English-learning infants able to distinguish between syllable-initial [m] and [n] but not between syllable-initial [n] and [ŋ] at 10-12 months, 6-8 months, and even at 4-5 months of age. Infants acquiring Filipino, which does contrast between syllable-initial [n] and [ŋ], do not show the ability to distinguish between syllable-initial [n] and [ŋ] at 6-8 months, only showing the ability to distinguish these sounds at 10-12 months of age. Taken together, these findings suggest that infants do not begin with an ability to distinguish *all* linguistically-relevant contrasts (in this case, [n-ŋ] or [d-ð]), but instead require experience to gain sensitivity to at least some contrasts.

Aslin and Pisoni (1980b) suggest a typology of possible developmental trajectories that contrasts may undergo during acquisition:

1. **Maintenance or facilitation.** Initially high or partially-developed sensitivity to a contrast remains high (“maintenance”) or improves (“facilitation”) with exposure.

*Example: The continued high sensitivity to syllable-initial [m-n] exhibited by English learning infants (Narayan et al., 2010).*

2. **Induction.** Initially poor sensitivity to a contrast improves with exposure.

*Example: English infants’ sensitivity to [d-ð] (Polka et al., 2001).*

3. **Loss.** Initially high or partially-developed sensitivity to a contrast declines from lack of exposure.

*Example: The decline in English infants’ sensitivity to [d-d] (Werker and Tees, 1984); the decline in Japanese infants’ sensitivity to [ɹ-l] (Iverson et al., 2003).*

4. **No effect.** Initially poor sensitivity remains poor with lack of exposure.

*Example: French infants’ sensitivity to [d-ð] (Polka et al., 2001; Sundara et al., 2006).*

To this list Cristia et al. (2011) add two types of developmental trajectories: poor sensitivity failing to improve with exposure, and high sensitivity remaining high with lack of exposure. It could be argued that non-sibilant fricatives fall into the first category, as Jongman et al. (2003) find that English speakers experience some difficulty distinguishing [f] and [θ] as well as [v] and [ð] despite these phones being contrastive in English; and it could also be argued that English speakers’ high sensitivity to clicks fall into

the second category, as Best et al. (1988) finds that English speaking adults and English-learning infants show high levels of discrimination to Zulu clicks despite clicks not falling within the English inventory and therefore not being linguistically contrastive.

To summarize, humans begin with natural discriminatory boundaries between some speech sounds (e.g. a VOT of around 20-50 ms for stops, the boundary between [m-n]). At least some of these boundaries are not human-specific and have been found, for example, in chinchillas and macaque monkeys. Humans also begin with no natural boundaries between other speech sounds (e.g. [d-ð], [n-ŋ]). To borrow an analogy from Cristia et al. (2011), infant perception begins as a topographical map, with natural peaks separating speech sounds. Through experience, this initial perceptual topography is warped such that new peaks are formed or existing peaks are flattened. The end result should be a language-specific perceptual map which aids in the discrimination of all contrastive phonemes in the target language.

This typology illustrates that language exposure influences sensitivities at least to some contrasts. But what aspect of the exposure results in this language-specific discrimination? One proposal is that infants learn which phones are contrastive in their language by learning minimal pairs (MacKain, 1982; MacKain and Stern, 1985). If a pair of words differing by exactly one sound have different meanings, infants can infer that the sounds which differ are contrastive in the language they are hearing, and therefore that the distinction between these sounds is an important one to make. Because of this, they can attend to whatever phonetic dimension(s) the two sounds differ by. However, this **minimal pair hypothesis** predicts that infants learn minimal pairs *before* exhibiting language-specific discrimination, something which does not appear to be the case. Language-specific discrimination appears around 6-12 months of age, but Caselli et al. (1995) finds that infants at 8 months of age have a receptive vocabulary of only about 36 words, none of which are minimal pairs. Further, a minimal pair hypothesis would predict that infants who are able to discriminate some contrast also attend to this difference in a word-learning task. Stager and Werker (1997) presented 14-month old infants with a picture and some label, [bɪ] or [dɪ]. After familiarization, experimenters tested infants' responses to the same label they had been trained on ([bɪ] or [dɪ]),

or on a minimally-different label in a switch trial ([dɪ] or [bɪ]). Stager and Werker found that these 14-month olds did not make use of their discriminatory ability to distinguish [b] and [d] in this switch task, suggesting that infants do not attend to phonetic detail in early word learning (also see Pater et al., 2004). Although further studies have found that infants in this age range can learn minimally-differing words in certain situations (Rost and McMurray, 2009; 2010; Galle et al., 2015; Fennell and Waxman, 2010), the minimal pair hypothesis would still predict that infants make use of phonetic differences during word learning, rather than ignore phonetic details in these switch tasks.

Another proposal for how this topography is warped is known as distributional learning (Maye et al., 2000; Maye and Gerken, 2002; Werker et al., 2012). According to this account, language learners map tokens into some phonetic space and make use of the relative frequencies at which tokens cluster in regions of this space to infer the number of phonetic categories in the language they are being exposed to (Maye and Gerken, 2002; Boersma et al., 2003; Guenther and Gjaja, 1996). Learners exposed to a bimodal distribution of tokens along some phonetic dimension(s) will infer that there are two phonetic categories, whereas learners exposed to a monomodal distribution will infer that there is only one phonetic category.

It is unclear whether distributions found in natural languages appear to support a distributional learning hypothesis. What does seem to be clear is that different phonetic categories exhibit different degrees of overlap with other phonetic categories (Moeng, 2016). Figure 5 shows the distribution of velar stops mapped along the dimension of VOT as measured by Lisker and Abramson (1964) for English. English has two velar stop phonemes, prevoiced velar stop /g/ and voiceless stop /k/. The voiceless stop /k/ has two allophones, [k<sup>h</sup>] found syllable-initially and [k] found elsewhere (Zsiga, 2013).<sup>4</sup> Based on the data collected by Lisker and Abramson, English speakers appear to be exposed to two large distribution

---

<sup>4</sup> For simplicity, other allophones such as unreleased [k̚] are ignored.



peaks, and one small one. One could imagine two models which explain this data in a way that is compatible with a distributional learning hypothesis. Either learners only notice distribution peaks which fall above some threshold, which would lead the English learner to notice only the two large distribution peaks shown in Figure 5, or the English learners notice all three peaks, but also learn that two of these peaks correspond to allophones of a single phoneme. Either way, the English speakers arrive at the conclusion that there are two phonemes, as predicted by distributional learning.

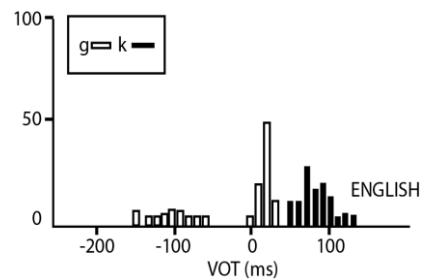


Figure 5. In a figure plotting VOT for velar oral stops, two or three clear peaks can be seen in the English-speaking data. Figure adapted from Lisker and Abramson (1964).

Although this data appears to fit well with a model of distributional learning, vowels have been noted to exhibit a high level of overlap with other vowel categories. Swingley (2009) maps 11 English monophthongs into an F1 vs. F2 space, as well as an F2-F1 vs. duration space. He finds a high level of overlap among these vowels for both of these spaces. A similar claim is made regarding vowel length in Japanese. Vowel length is contrastive in Japanese, so a theory of distributional learning would predict that two peaks appear for each of the five vowel qualities in Japanese. Bion et al. (2013) measure length for naturally-produced vowels in infant-directed speech. Although they find a significant difference in length between long and short vowels, Bion and colleagues fail to find clear peaks in distribution along the duration dimension, especially for the vowel qualities [a], [e], and [u].

The question of whether natural language provides infants with distributions which are clear enough to learn from is complicated by the fact that it is unclear how many dimensions a language learner might make use of when mapping tokens in some space. For example, Bion et al. (2013) only map Japanese vowels along one dimension, vowel duration, but it is possible that Japanese long and short vowels

differ in some other acoustic factors other than length, and that mapping these vowels along multiple dimensions may have resulted in a clearer distinction between long and short vowels in Japanese. Swingley (2009) considers an F1 vs. F2 space when mapping English monophthongs, as well as an F2-F1 vs. duration space, but in reality it is unknown whether language learners make use of two dimensions or twenty.

However, if we put aside the issue of dimensionality and assume that overlap is problematic for at least some phonetic categories, various researchers have suggested supplementary cues that infants might make use of. Adriaans and Swingley (2012) suggest that infant-directed speech aids infants in finding peaks in distribution by directing infants through prosody to “high quality” tokens. Infants can then treat these “high quality” tokens as being more important when determining phonetic categories. Adriaans and Swingley show that mapping only “focused” vowel tokens, those which are prosodically-exaggerated (either through a longer duration, higher average pitch, and/or larger change in pitch compared to the average vowel) reduces the level of overlap exhibited in English vowels compared to mapping all vowel tokens, exaggerated or not. Others have proposed bootstrapping methods learners might use, such as overall wordform (Feldman et al., 2009; 2011, 2013; Thiessen, 2007) or knowledge of very common words (Swingley, 2009, 2007), enabling a theory of phonetic category acquisition that is not solely dependent on phone distributions. Feldman et al. (2011) find that presenting learners with sounds in different lexical environments serves to distinguish those sounds. For example, learners exposed to [guta] and [lito], but not [lita] and [guto], will have a greater sensitivity to [ta] and [to] following training compared to learners exposed to [guta], [guto], [lita] and [lito]. This has been found for adult learners (Feldman et al., 2011) and infants (Feldman et al., 2013; Thiessen, 2007); for an [a-ɔ] contrast (Feldman et al., 2011, 2013), as well as a [t-d] contrast (Thiessen, 2007).

As will be shown below, regardless of whether or not natural language exhibits clear distributions, both adults and infants *are* able to make use of clear distributional information in a lab setting, lending support to a distributional learning mechanism.

## 2.2. EXPERIMENTAL SUPPORT

Maye and Gerken (2000) is the first artificial language learning study which provides experimental support for distributional learning. In this study, adult participants were exposed to CV syllables during a training phase which lasted 9 minutes. Exposure syllables during this training phase consisted of fillers, [mɑ mæ mə lɑ læ lə], as well as three 8-point continua: 8 tokens ranging between [dɑ] and [ɔɑ], 8 tokens ranging between [dæ] and [ɔæ], and 8 tokens ranging between [də] and [ɔə]. Continua were created by first recording a native English speaker producing the syllables [dɑ dæ də] and [stɑ stæ stə]. The initial [s] was then removed from the latter three syllables. Three 8-point continua, one for each vowel context, were created through re-synthesis.<sup>5</sup> The training phase consisted of 384 syllables. Participants were given a check sheet with 384 blank boxes on it and were instructed to simply listen and check one box for each syllable they heard during this training phase.

Participants were randomly assigned to one of two groups: a Bimodal group and a Monomodal group. These two groups only differed in the frequency distributions that critical stimuli were presented in. Participants in the Monomodal group were exposed to a monomodal distribution of continuum points, such that continuum points near the center of the continuum (points 4 and 5), were presented four times as frequently as continuum points near the endpoints of the continuum (points 1, 2, 7, and 8) (see solid line in Figure 15). Participants in the Bimodal group were exposed to a bimodal distribution of continuum points, such that continuum points near the endpoints (points 2 and 7) were presented four times as frequently as continuum points at the endpoints (points 1 and 8), and continuum points at the center of the continuum (points 4 and 5) (see dotted line in Figure 15). Continuum points will be referred to here as D<sub>1</sub>-D<sub>8</sub>.

---

<sup>5</sup> Further details were not given.

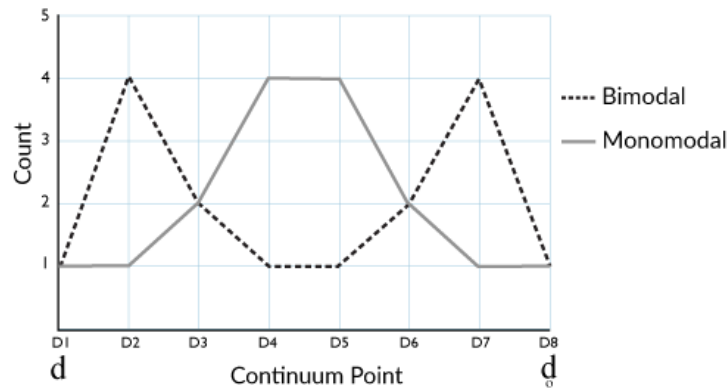


Figure 6. Illustration of familiarization frequency of onsets of critical stimuli for Bimodal and Monomodal groups during the training phase of Maye and Gerken (2000).

Following training, participants were directed to a test phase. In the test phase, participants were presented with pairs of syllables separated by 500 ms and were asked if they believed the two syllables presented were the same word repeated twice, or two different words, in the language they had been exposed to. During this phase, participants heard four types of syllable pairs: filler Same Pairs, which consisted of non-identical tokens of the same filler syllable (e.g. [mæ]<sub>1</sub> vs. [mæ]<sub>2</sub>); filler Different Pairs, which consisted of different filler syllables (e.g. [mæ] vs. [læ]); critical Same Pairs, which consisted of identical tokens of critical syllables taken from the same end of the continuum (e.g. D<sub>1</sub>æ vs. D<sub>1</sub>æ); and critical Different Pairs, which consisted of tokens taken from opposite ends of the continuum (e.g. D<sub>1</sub>æ vs. D<sub>8</sub>æ). Maye and Gerken found a greater percentage of “different” responses for critical Different Pairs in the Bimodal condition than in the Monomodal condition. They concluded that this study supports the theory of distributional learning.

Maye and colleague’s findings are widely cited (for example, Kuhl, 2004; Kuhl et al., 2006; Werker et al., 2012), and have been replicated a number of times. Experimental support has been found for adults (Maye and Gerken, 2000; Maye and Gerken, 2001; Hayes-Harb, 2007; Escudero et al., 2011) and infants (Maye et al., 2002; 2008; see Cristia, 2018 for a meta-analysis). Attempts to replicate Maye and Gerken’s (2000) findings with other stimuli have shown mixed success. Stimuli successfully used in replications include the stop pairs [d-d̥], and [g-g̥] (Maye and Gerken, 2000; Maye and Gerken, 2001;

Maye et al., 2002; Hayes-Harb, 2007); the vowel pairs [a-ɑ], and [i-i] (Wanrooij et al., 2013; Gulian et al., 2007; Escudero et al., 2011; Escudero and Williams, 2014); and the Thai tone pairs [33] and [241] (Ong et al., 2015). However, Peperkamp et al. (2003) failed to replicate these findings when testing fricatives ranging from [ʁ] to [χ] with French-speaking adult participants. And although the Dutch contrast between [a] and [ɑ] has been successfully tested in distributional learning studies with adult speakers of Spanish (Wanrooij et al., 2013; Escudero et al., 2011), Ong et al. (2016) failed to replicate these findings when testing the same contrast with Australian English-speaking adult participants. The authors attribute this lack of replication to the higher initial discriminatory ability of [a] and [ɑ] by Australian English speakers compared to Spanish speakers, which may indicate that distributional learning can only increase sensitivity and not decrease it. Maye and Gerken (2001) find that distributional learning of one acoustic cue (e.g. [d] vs. [d̥]) fails to extend to a new contrast which varies along the same dimension (e.g. [g] vs. [g̥]), but Perfors and Dunbar (2010) find that participants can extend distributional learning to a new contrast if training is intensified (i.e. is longer and contains no fillers). Escudero and Williams (2014) find evidence that distributional training on adults has long-term effects (up to 12 months) on discriminatory abilities. In a meta-analysis, Cristia (2018) concludes that distributional learning as studied on infants using a habituation/change design shows a robust effect, but that distributional learning as studied with an alternating/non-alternating design does not. A further description of her meta-analysis will be given in Chapter 3.

### 2.3. MODELS OF DISTRIBUTIONAL LEARNING

Previous explanations of distributional learning rely on acoustic input warping a listener's perceptual space (Boersma et al., 2003; Guenther and Gjaja, 1996). Guenther and Gjaja (1996) model distributional learning with self-organizing neural networks. Their model consists of two layers of cells: formant representation cells and auditory map cells. Each formant representation cell is connected to all auditory map cells through synapses. Depending on input from formant representation cells and on synapse strength, a subset of auditory map cells is activated. Activated cells determine what sound is perceived. Learning in

this model consists of modifying synapse strengths, so that frequent auditory tokens lead to strengthened firing preferences of auditory map cells, and infrequent auditory tokens lead to weaker firing preferences. This leads to a perceptual warping of the space such that two auditory tokens near category centers are perceived as being more similar to one another compared to two equally-spaced auditory tokens that are located further away from the category center.

Boersma et al. (2003) suggest another perceptual warping model of distributional learning in a constraint-based model. In their model, the input consists of an auditory value (e.g. [F1: 300 Hz]), and the output consists of a perceived phonetic category (e.g. /F1: 320 Hz/). In this Optimality Theoretic model, there are three families of constraints responsible for distributional learning: \*CATEG, PERCEIVE, and \*WARP. This model will be further described in Chapter 3.

Both the Guenther and Gjaja (1996) model and the Boersma et al. (2003) model are similar in that (1) uneven frequency distributions along some acoustic dimension(s) lead to uneven mappings between auditory input and perceived value, thereby warping the listener's perceptual space, and (2) learners are not required to hold a large number of exemplars in memory. Rather, experience changes some aspect of the entire perceptual system (i.e. through constraint rankings or synapse strengths).

Chapter 3 will discuss two terms, **response bias** and **sensitivity**, and argue that models of perceptual warping such as those described by Guenther and Gjaja (1996) and Boersma et al. (2003) predict that distributional learning should always be accompanied by a change in sensitivity. The distinction between bias and sensitivity comes from ideas in Signal Detection Theory, which is outlined in the following section.

### **3. Main Concepts in Signal Detection Theory**

This study makes a distinction between **response bias** (simply, “bias”) and **sensitivity** in distributional learning. These concepts are used in Signal Detection Theory (Peterson et al., 1954; Swets, 2014), which models the perceiver's internal process of discriminating whether some signal is or is not present. Signal

Detection Theory assumes that there is some inherent amount of uncertainty in a discrimination process which comes in the form of *internal noise*. The concept of internal noise is based in the idea that whatever *internal response* is occurring during the decision-making process, such as neurons firing in the brain, is noisy by nature. In Signal Detection Theory, the decision-making process is assumed to be based on this internal response. An organism's internal response can be modelled as falling along some numeric line, with larger values indicating a greater internal response. The greater the internal response, the greater the probability that an organism perceives some signal (whether the signal is actually present or not). Due to the inherent internal noise of an organism's neural response, the probability curve of an organism's internal response is assumed to be normally distributed.

As an example, suppose we are modelling the detection process of a participant taking part in a lexical decision task. The participant is presented with both Real Word and Nonce Word trials at random, in which a real or a nonce word is played to the participant. The participant is asked to press a 'yes' button if the word he or she hears is a real word in English, and a 'no' button if it is not an English word. Figure 7 illustrates two internal response probability curves. The curve on the left represents the probability of the internal response experienced by the participant during a NonceWord trial (No Signal), while the curve on the right represents the probability of the internal response experienced by the participant during a RealWord trial (Signal). Note that both probability curves are normally (Gaussian) distributed, representing the inherent noise experienced by all organisms during neural activity. Overall though, the curve representing a NonceWord trial has a lower mean internal response than the curve representing a RealWord trial.

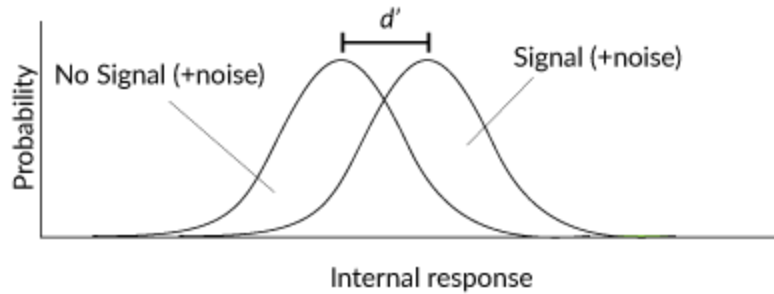


Figure 7. Internal response when there is a signal, represented by the curve on the right, compared to when there is no signal, represented by the curve on the left. In the example here, the Signal curve represents the probability of the participant's internal response during a RealWord trial, and the No Signal curve represents the probability of the participant's internal response during a NonceWord trial.

In Signal Detection Theory, sensitivity is represented by the distance between an organism's internal response when there is a signal and when there is no signal. This distance is called  $d'$  (*d-prime*) and is shown as the distance between the means of both probability curves. A greater  $d'$  represents the internal state of an organism that is more able to distinguish between a Signal and No Signal situation. Compare the distance between the Signal and No Signal probability curves in Figure 7 and Figure 8. Figure 8 represents the internal state of an organism with a higher sensitivity (or, a greater  $d'$ ) to a signal than that shown in Figure 7, since the organism's internal response when there is a signal is more distinct from when there is no signal. If  $d'$  is zero, then the organism is unable to distinguish when there is and is not a signal since the Signal and No Signal probability curves would fall in identical locations.

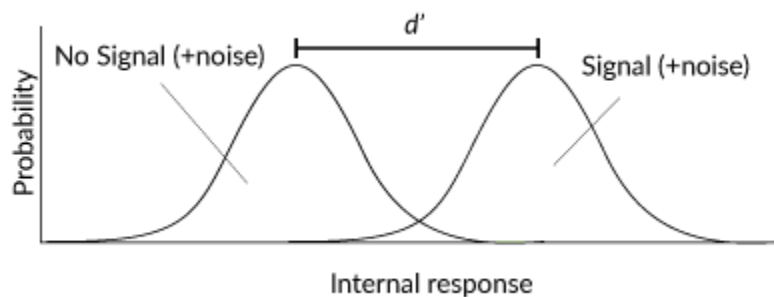


Figure 8. Internal state for a signal to which the organism has great sensitivity to, where sensitivity is represented by  $d'$ .

In order to make a decision, the participant has some *criterion response* level. Returning to our lexical decision task example, if the internal response experienced is greater than the criterion that this particular



participant has, he or she will press the ‘yes’ button, represented by the light-grey and light-striped areas in Figure 9. If the internal response experienced is less than the criterion, he or she will press the ‘no’ button, represented by the dark-grey and dark-striped areas in Figure 9.

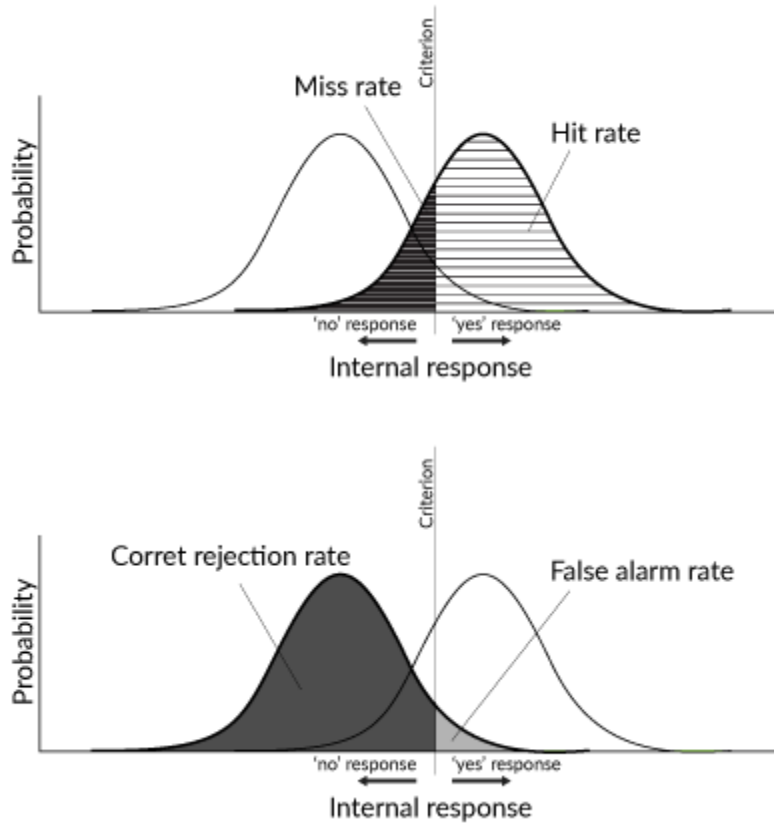


Figure 9. Criterion response defines four probabilities: for RealWord trials (top) the location of a participant’s criterion defines the participant’s hit rate and miss rate; for NonceWord trials (bottom) the location of a participant’s criterion defines the participant’s false alarm and correct rejection rate.

Therefore, since the light-striped area comes from the probability curve representing the participant’s internal response during a RealWord trial, the light-striped area represents the probability that the participant correctly responds ‘yes’ when there is a signal. This is called the participant’s *hit rate*. The light-grey area comes from the probability curve representing the participant’s internal response during a NonceWord trial, so the light-grey area represents the probability that the participant incorrectly responds

‘yes’ when there is no signal. This is called the participant’s *false alarm rate*. The dark-striped area represents the probability that the participant incorrectly responded ‘no’ when there was a signal (*miss rate*), and the dark-grey area represents the probability that the participant correctly responded ‘no’ when there was no signal (*correct rejection rate*). These probabilities are summarized in Figure 10.

		Signal	
		Present (e.g. RealWord trial)	Absent (e.g. NonceWord trial)
Response	Present (e.g. ‘yes’ response)	Hit	False alarm
	Absent (e.g. ‘no’ response)	Miss	Correct rejection

Figure 10. Participant responses can be divided into four categories: hits, false alarms, misses, and correct rejections. In the example given here, a present Signal is represented by a RealWord trial, and an absent Signal is represented by a NonceWord trial. A ‘present’ response is represented by a participant responding ‘yes,’ and an ‘absent’ response is represented by a participant responding ‘no.’

Note that in Signal Detection Theory, a participant’s sensitivity to stimuli  $d'$  (represented by the distance between a Signal and No Signal curve) is independent of the criterion established by the participant (Macmillan and Creelman, 2004:36; also see Stanislaw and Todorov (1999) for examples). Some participants may be more inclined to respond ‘yes,’ while others are more conservative in their responses and inclined to respond ‘no.’

Figure 11 represents two participants’ internal states. The top figure shows the internal state of a participant who has established a high criterion, and therefore responds ‘no’ most of the time. The bottom figure shows the internal state of a participant who has established a low criterion, responding ‘yes’ most of the time. Note that the change in criterion levels does not affect sensitivity: both of these participants have the same sensitivity  $d'$ .

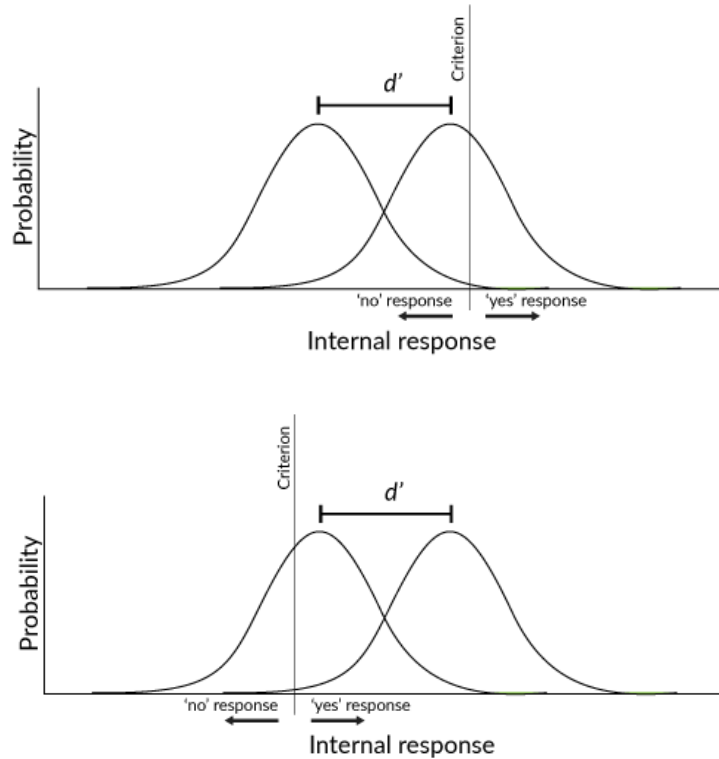


Figure 11. Internal state of a participant with a high criterion (top figure) compared to internal state of a participant with a low criterion (bottom figure). Note that these particular hypothetical participants have identical sensitivities.

Signal Detection Theory assumes that internal responses are normal in their distribution. Because of this, a participant's sensitivity to stimuli can be calculated if their hit and false alarm rates are known. The z-score of the probability that a participant correctly responds 'yes' when a signal is present (hit rate) will give us the number of standard deviations that the criterion is from the Signal probability curve mean. The z-score of the probability that a participant *incorrectly* responds 'yes' when a signal is absent (false alarm rate) will give us the number of standard deviations that the criterion is from the No Signal probability curve mean. If we then subtract the z-score of the false alarm rate from the z-score of the hit rate, this gives us the distance between the Signal and No Signal curve (Figure 12).

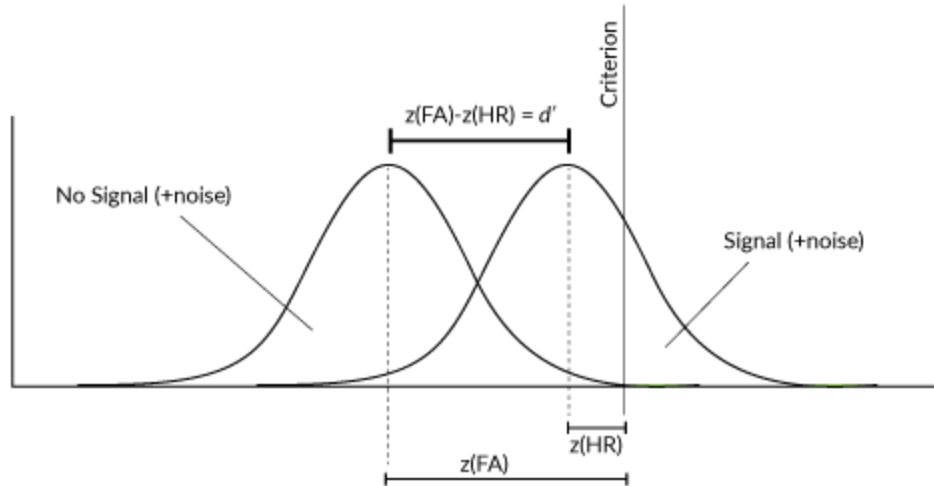


Figure 12. Calculation of  $d'$ .

To summarize, a participant's sensitivity can be calculated using the equation in (1).

$$(1) \quad d' = z(FA) - z(HR)$$

Past distributional learning studies which have worked within a Signal Detection framework have calculated  $d'$  values for individual participants, and then analyzed those values using an ANOVA (e.g. McGuire, 2007; Hayes-Harb, 2007; Noguchi, 2016). This dissertation does not directly calculate  $d'$ , but instead analyzes data using a generalized linear mixed effects model with a logistic link function, which is formally identical to Choice Theory, a close cousin of Signal Detection Theory. This is done for two reasons: 1) ANOVAs assume data is normally distributed (which data presented in this dissertation is not<sup>6</sup>), and 2) logistic regressions have been argued to be superior in analyzing categorical data (see Jaeger, 2008). The following section describes the relationship between Choice Theory and logistic regressions.

---

<sup>6</sup> This is especially the case in Experiments A1 and A2, in which responses are heavily skewed towards “same” responses.

### 3.1. THE RELATIONSHIP BETWEEN CHOICE THEORY AND LOGISTIC REGRESSIONS

Choice Theory (Luce, 1959) is formally similar to Signal Detection Theory, with the exception that Choice Theory assumes that internal noise distributions are logistic rather than normal (Macmillan and Creelman, 2004). DeCarlo (1998) further shows that Choice Theory and logistic regressions are formally identical, allowing for the statistical analysis of sensitivity and response bias using logistic regressions.

Several questions arise in using logistic regressions to analyze sensitivity and response bias:

- 1) How does one interpret sensitivity in a logistic regression?
- 2) How does one interpret response bias in a logistic regression?
- 3) How should data be interpreted when both sensitivity and response bias are affected?
- 4) How should *same-different* experiments be analyzed in Choice Theory?

The remainder of this subsection will attempt to provide answers for these questions and clarify which simplifying assumptions will be made in this dissertation.

### 3.2. SENSITIVITY IN LOGISTIC REGRESSIONS

To consider how sensitivity should be interpreted in a logistic regression, it may be useful to review the figures from the previous section below.

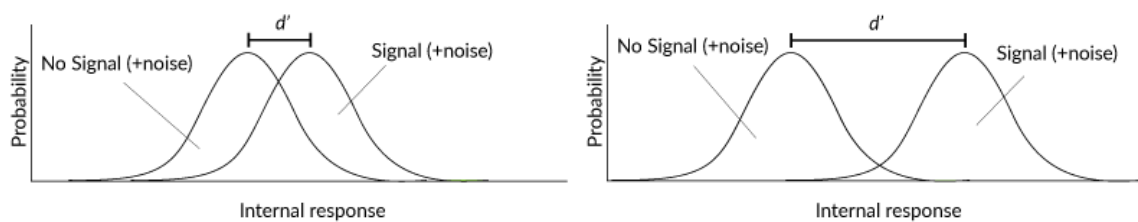


Figure 13. Figure on the right shows increased sensitivity to Signal and No Signal stimuli compared to the figure on the left.

Figure 13 shows that sensitivity, or  $d'$ , measures the distance between the mean internal responses for No Signal and Signal trials. Suppose that the figure on the left illustrates participants in Condition X, while the figure on the right illustrates participants in Condition Y. In terms of a logistic regression, the effect of

one factor, stimulus type (Signal or No Signal), depends on the level of another factor, i.e. which condition participants are in. In short, Condition X and Condition Y have different sensitivities if one finds a significant interaction between stimulus type (Signal or No Signal) and condition (see Macmillan and Creelman, 2004; DeCarlo, 1998). Therefore, this dissertation will interpret a significant interaction between stimulus type and condition as a difference in sensitivity between conditions, where the condition with the greater sensitivity has a greater distance between mean responses between Signal and No Signal stimuli.

### 3.3. RESPONSE BIAS IN LOGISTIC REGRESSIONS

To consider bias should be interpreted in a logistic regression, consider again the figures below.

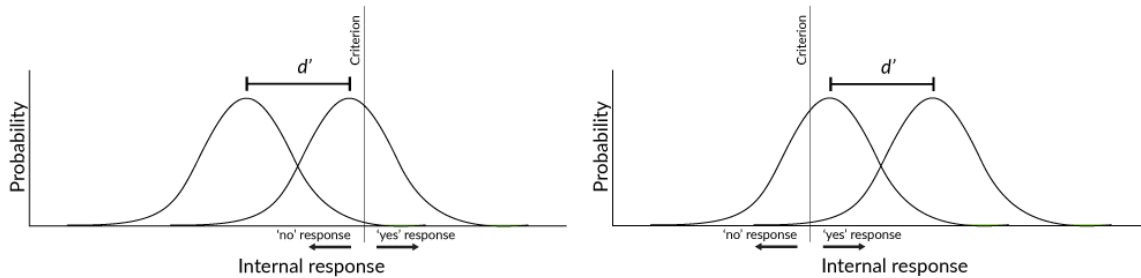


Figure 14. Figure on the right shows increased bias towards “yes” responses compared to the figure on the left.

Figure 14 shows two situations with differing response criteria, but the same sensitivity, or  $d'$ . Again, suppose that the figure on the left illustrates participants in Condition X, while the figure on the right illustrates participants in Condition Y. In terms of a logistic regression, the odds of a “yes” response increases for all stimuli type, both Signal and No Signal stimuli alike, between Condition X and Condition Y. That is, the factor of condition has an effect on response, regardless of stimulus type. This case would present itself as a main effect of condition (see Macmillan and Creelman, 2004; DeCarlo, 1998). Therefore this dissertation will interpret a significant main effect of condition as a difference in response bias between conditions, where the condition with the greater mean “different” responses has a greater bias towards a “different” response.

### 3.4. SENSITIVITY AND RESPONSE BIAS IN LOGISTIC REGRESSIONS

Suppose a situation where a significant interaction between condition and stimulus type is found, and a significant main effect of condition is also found. Would this mean that Condition X and Condition Y differ in both sensitivity and response bias? This subsection argues that this is not always an accurate inference that can be made.

If an interaction between condition and stimulus type is found, then the effect of stimulus type on response is dependent on condition. If this is the case, it does not make sense to average response over levels of the stimulus type factor, since it has already been found that the effect of each of the levels of stimulus type depends on condition. In other words, if we find that Condition Y has increased sensitivity compared to Condition X, and also find that Condition Y is more likely to respond “yes” when we average over all stimulus types, we cannot be certain that this overall increased “yes” response comes (1) solely from one level of the stimulus type factor, (2) from all levels of the stimulus type factor, or (3) from some combination of (1) and (2). The reason we cannot disambiguate between each of these three situations is that we already know that the effect of condition depends on stimulus type through the significant interaction between condition and stimulus type. Therefore, in the same way main effects are not interpretable when there is an interaction, we cannot determine through main effects whether response bias was also affected or not if we also find a significant difference in sensitivities between conditions. For the purposes of simplification, this dissertation will not attempt to disambiguate situations in which both a significant interaction and a significant main effect are found. Therefore, this dissertation will not interpret main effects as a difference in response bias between conditions if an interaction between condition and stimulus type is found.

### 3.5. ADAPTING CHOICE THEORY TO SAME-DIFFERENT EXPERIMENTS

In addition to the above simplification, this dissertation will also make the assumption that participants treat the difference in pairs of sounds played side by side as a Signal or No Signal stimulus. That is, this dissertation assumes that the “signal” being detected is a difference in acoustics. This is not the only way

to analyze *same-different* experiments through signal detection type models, and DeCarlo (2013) argues for several alternate analyses using nonlinear models. However, for the purposes of this dissertation, Different Pairs will simply be treated as “Signals,” and Same Pairs will be treated as “No Signal” stimuli so that data can be analyzed using generalized linear models.

### 3.6. SUMMARY OF ANALYSIS TO BE USED

To briefly summarize, this dissertation will adhere to the following conventions to interpret findings:

- (1) A significant interaction between condition and stimulus type will be interpreted as a significant difference in **sensitivity** between conditions.
- (2) A significant main effect of condition will be interpreted as a significant difference in **response bias** between conditions...
- (3) ... *unless* a significant interaction between condition and stimulus type was also found, in which case a main effect will not be interpreted.

Additionally, *same-different* experiments will be treated as *no-yes* or *noSignal-signal* experiments.

## 4. Variations in Distributional Learning Experiments

This section returns to distributional learning, with the aim of providing a typology of past methodologies used in studies which have been modelled on Maye and Gerken (2000). I believe this is important as small changes in both methodology and analysis can potentially result in the measurement of different aspects of distributional learning.

Maye and Gerken (2000) were originally interested in participants’ decision-making, categorization skills. They were interested in determining whether, given two perceptibly-different syllables, participants categorized these tokens as belonging to the same category, or to two separate categories. This would be similar to being given a dark blue swatch and a light blue swatch and being asked whether these belonged to the “same” category. Viewers are able to distinguish the two, but would say they belong to the “same” category since they are both “blue.” This type of task tests participants’ upper-level decision-making process, rather than any low-level discriminatory process. I will refer to this as an “open-numbered categorization.”



This is opposed to a discrimination task, where participants hear two very similar syllables, and must determine whether the two are identical tokens, or if they are acoustically different from one another. This would be similar to looking at two very identical light blue color swatches and determining if there is any perceptible difference between the two.

Maye and Gerken (2000) used a *same-different* task to analyze participants' categorizations. In order to ensure that they were analyzing categorization rather than discrimination, they included non-identical filler tokens. "Same" fillers were different recordings of the "same" syllable (e.g. two separate recordings of the syllable *ma*). In this way, the experimenters meant to ensure that participants were basing their responses on upper-level decision-making ("open-numbered categorization"), and were not simply listening for any acoustic difference ("discrimination"). Despite this safeguard, when listening to these tokens myself, the "same" fillers did indeed sound identical to my ears. Even if these tokens are different recordings, if they fall below the level of perceptible difference to a human listener (as they did for me), they do not necessarily keep the participant from treating the task as a simple discrimination task. Because of this I believe the *same-different* task used by Maye and Gerken is still ambiguous between being an *open-numbered categorization* task (e.g. participants respond that dark blue and light blue are the "same" since they are both blue, despite being perceptibly distinct from one another) and a *discrimination* task (e.g. participants respond that two swatches of blue are only the "same" if they are perceptibly identical).

Rather than using a *same-different* methodology, some distributional studies have opted to use an *ABX* (or *XAB*) task to test participants (Ong et al., 2015). In this task, participants hear three syllables, and are asked to categorize the third syllable as belonging to the category of the first syllable or the second syllable (or the first syllable as belonging to the category of the second or third syllable in an *XAB* task). If it is the case that *X* is identical to *A* or to *B* (or close enough that a participant perceives the two as being identical), this task becomes a discrimination task. If all three syllables are acoustically distinct

enough from one another to be discriminable by the participant, this task would be a “closed-number categorization task.” That is, the nature of the task is such that the participant knows that there are two categories, *A* and *B*, among the three sounds they are hearing. Their task is simply to determine whether *X* belongs to the category represented by *A*, or the category represented by *B*. (This section makes a distinction between *closed-number* and *open-numbered categorization* tasks, because, as will be discussed in further chapters, I believe that distributional learning begins with a Bias stage, which would only be captured with an *open-numbered categorization* task.)

Since the stimuli used in a study are often not available for future readers to access, it is difficult to categorize past studies as being *open-numbered categorization* tasks, *closed-number categorization* tasks, or *discrimination* tasks. The studies presented in this dissertation will follow Maye and Gerken (2000) in using a *same-different* methodology with Same Pair fillers which are discriminable from one another, in hopes that participants will treat the task as an *open-numbered categorization* task.

Past studies also vary in the methods used to analyze participants’ responses. Some studies measure the percentage of “correct” responses in Different Pairs (Maye and Gerken, 2001; Pajak and Levy, 2011; Hayes-Harb, 2007), while others measure *d-prime* (Noguchi, 2016; Hayes-Harb, 2007). As will be further discussed in Chapter 3, I believe these two measure different aspects of learners’ responses.

So far, this chapter has blurred the distinction between first language acquisition, second language acquisition, and artificial language learning experiments. The theory of distributional learning is motivated by infants’ developmental trajectories. That is, infants exhibit language-specific perceptual warping from 6-12 months of age, before they know enough minimal pairs for this warping to be attributed solely to lexical learning (although see Swingley (2009) and Feldman et al. (2013) for supplemental learning based on lexical form). Yet, several experiments supporting distributional learning have been conducted on adults. Because this dissertation bases its proposal on results from artificial language learning experiments conducted on adults, this inconsistency is addressed in the following section

by providing justification for using artificial language experiments to study distributional learning, but also by acknowledging the need for replication work with infants.

## **5. The Relationship Between Natural and Artificial Language Learning**

Artificial language learning studies offer us a unique window into human language acquisition. However, as with any methodology, these types of studies do have weaknesses that should be acknowledged. This section provides justification for studying distributional learning with artificial language learning experiments on adults, while also acknowledging weaknesses of this methodology.

One of the main weaknesses of artificial language learning studies has to do with the unclear status as to what is being modelled. On one hand, it could be argued that artificial languages are indicative of second language acquisition, since adult participants experience interference from their first language (Schwartz and Sprouse, 1996) and have passed any theorized critical period of language acquisition (Lenneberg, 1967). Even if one assumes that various mechanisms used in language learning function throughout a language learner's life, it is possible that some mechanisms are stronger at various points in development. For example, Thiessen and Saffran (2003) tested 7-month and 9-month infants in a speech segmentation task. Infants were presented with conflicting cues as to where word boundaries occur. The 9-month old infants used stress to segment speech into words when stress and statistical cues indicated different word boundaries, tending to segment speech into trochaic "words," while the 7-month old infants used statistical cues, tending to segment speech at regions of low transitional probabilities. Thiessen and Saffran suggest that infants rely more heavily on statistical cues early on (at 7 months), and only make use of other cues, such as typical stress patterns, later on (at 9 months). If differences in cue weighting are apparent at 7 months compared to 9 months of age, there must be a number of differences between cue weightings used by adults and those used by infants. On the other hand, one could argue that artificial language learning studies are reflective of first language acquisition, since (although dependent on the particular study's methodology) artificial language learning studies often do not present learners with the type of explicit instructions that a second language learner might receive (for example, Maye and

Gerken, 2000; Saffran et al., 1997; Noguchi, 2016; Feldman et al., 2011), which can have an effect on acquisition (see Zhang (2013) for an example of the effect of explicit instruction on Chinese tone acquisition). Some artificial language learning studies draw conclusions about the nature of second language acquisition (Hayes-Harb, 2007; Escudero et al., 2011), while others draw tentative conclusions about the nature of first language acquisition (Peperkamp et al., 2003; Maye and Gerken, 2001; Noguchi, 2016). A number of adult studies are followed up with a replication study on infants before conclusions regarding first language acquisition are made (e.g. Maye and Gerken (2000) followed by Maye et al. (2002; 2008); Feldman et al. (2011) followed by Feldman et al. (2013)). There is also the very real possibility that artificial languages are not indicative of any natural linguistic process (for a discussion of this possibility, see Moreton and Pater, 2012). The ambiguity in what is being modelled in adult artificial language learning studies is particularly important to keep in mind since the main topic under investigation here, distributional learning, was primarily motivated by observations of language development in infants (Maye and Gerken, 2000), as summarized in the previous section.

With that being said, this dissertation presents a series of artificial language learning experiments conducted on adults. Since no experiments reported here are conducted on infants, this dissertation remains agnostic as to whether findings reflect general cognitive mechanisms utilized by both adults and infants in language acquisition, or are only indicative of adult cognition, and will simply refer to “language learners.” It is still hoped that results can be extended to first language acquisition, since the overall topic under study, distributional learning, has been found in both adult (Maye and Gerken, 2000; Ong et al., 2015; Hayes-Harb, 2007; Maye and Gerken, 2001; Noguchi, 2016) and infant experiments (Maye et al., 2002; Yoshida et al., 2010), and since other phenomena also reliant on statistical tracking have been claimed to occur in both adults and children (Saffran et al., 1997). When available, results from infant studies which appear to corroborate this dissertation’s findings will be mentioned.

## **6. Research Questions and Structure of Dissertation**

This dissertation is interested in defining a timeline of early phonological acquisition. Although initially meant as a simple replication study, the data presented in Chapter 3 suggests a necessary distinction between two stages of phonetic category acquisition, a Bias Stage and a Sensitivity Stage. Overall, this dissertation seeks to answer the following question:

- (1) What are the stages in early phonological acquisition?

The main goal of this dissertation is to formulate a timeline of phonological acquisition based on experimental evidence. In answering this question, this dissertation explores the interaction of distributional learning with various phenomena, indicated below:

- (2) How does distributional learning interact with attention? (Chapter 3)
- (3) How does distributional learning interact with environmental context? (Chapter 4)
- (4) How does distributional learning interact with word learning? (Chapter 5)

I argue in this dissertation that phonetic category learning occurs in two stages, a Bias Stage and a Sensitivity Stage. This will be supported by a series of experiments, the “A” Experiments, in Chapter 3. Results of the “A” Experiments also indicate an effect of attention on distributional learning. Chapter 4 explores the relationship between phonetic category acquisition and allophony acquisition, and presents experimental support for a one-stage model of allophony acquisition (Experiment B), as suggested by Dillon et al. (2013). Finally, this dissertation discusses the gap between phonetic category acquisition and functional phonemes that are used to differentiate words in a set of “C” Experiments, presented in Chapter 5.

## **7. Summary**

To summarize, this chapter showed that despite supplemental theories which have been suggested (Feldman et al., 2013; Thiessen, 2007), distributional learning makes up a large portion of the explanation of how infants come to exhibit language-specific discrimination at such an early age. Experimental support for distributional learning is primarily based on artificial language studies which have been conducted on

adults; but, importantly, these experiments have also been replicated on infants. This chapter then introduced two main concepts from Signal Detection Theory, *response bias* and *sensitivity*, and illustrated how these two concepts are independent of one another. Specific guidelines regarding the interpretation of this dissertation's experimental results were then laid out. Following this, this chapter highlighted seemingly small differences in both methodology and analysis between past distributional learning experiments. The next chapter will further discuss response bias and sensitivity. It will be shown that 1) the seemingly small differences in methodology and analysis of past distributional learning experiments have resulted in the measurement of different things, and 2) past models of distributional learning predict that sensitivity, and not necessarily bias, is directly affected by distributional learning. A set of experiments (the "A" Experiments) will be presented which counter these sensitivity models of distributional learning.

## Chapter 3:

### Response Bias and Sensitivity in Distributional Learning, and the Role of Attention

#### 1. Introduction

The original goal of this set of experiments was to determine whether distributional learning, which has been found in in-person experiments, could be replicated through an online platform. In the process of doing so, two main theoretical conclusions were also reached: (1) the distribution of exposure can affect learners' **response bias** (also simply “bias”) independently of their **sensitivity**, and (2) attention plays a role in distributional learning. The first contribution regarding the distinction between response bias and sensitivity in distributional learning is supported with three main “A” Experiments: Experiments A1, A2, and A3. The second theoretical contribution regarding attention in distributional learning will be discussed in two follow-up “Tone” experiments, A2-Tone and A3-Tone. Finally, as all experiments were conducted online using the online participant pool known as Mechanical Turk, this chapter ends with two more contributions which are methodological in nature, by giving suggestions for conducting perceptual experiments online.

Of these four contributions, the primary theoretical contribution of this chapter regards the distinction between response bias and sensitivity. This study argues for a **two-stage model of phonetic category acquisition**<sup>7</sup>: a Bias Stage followed by a Sensitivity Stage. This runs contrary to models which base distributional learning in perceptual warping (Guenther and Gjaja, 1996; Boersma et al., 2003), which do not predict a Bias Stage of phonetic category acquisition. Although all experiments conducted

---

<sup>7</sup> Not to be confused with the theoretical two-stage model of *allophonic* acquisition, discussed in Chapter 4.

here were conducted on adults, Section 7 will present evidence that this two-stage model is supported by past infant studies of distributional learning.

Section 2 provides background for this study. Section 3 states this chapter's main research questions and provides a summary of results from all experiments conducted in this chapter. Sections 4-6 describe the methodology and results of the three main "A" Experiments conducted, A1, A2, and A3. This is followed by a summary and discussion of response bias and sensitivity in distributional learning in Section 7, including possible supporting evidence of a distinction between bias and sensitivity from past distributional learning studies conducted on infants (Yoshida et al., 2010; Maye and Gerken, 2002). Section 8 presents two "Tone" experiments, Experiment A2-Tone and Experiment A3-Tone. Results of these experiments will be used to argue that **attention** plays a role in distributional learning in Section 9. Section 10 provides suggestions for those wishing to conduct perceptual experiments online through platforms such as Mechanical Turk. Section 11 discusses unexpected results involving filler trials and provides possible explanations for these results. Section 12 concludes with a summary of results and an overview of contributions made in this chapter.

## **2. Background**

This section begins with a summary of the methodology and conclusions of Maye and Gerken (2000) (Section 2.1), which forms the basis of the experiments conducted here. Following this, Section 2.2 provides a background on the medium used in these distributional learning experiments, Mechanical Turk. Section 2.3 describes the distinction between a learner's **response bias** (i.e. their inclination to respond one way or the other) and their **sensitivity** (i.e. their ability to perceive a phonetic distinction). Section 2.4 then simulates one perceptual warping model of distributional learning in order to show that this model predicts a change in sensitivity, but not in response bias.



## 2.1. DISTRIBUTIONAL LEARNING AND MAYE AND GERKEN (2000)

Distributional learning refers to a type of statistical learning in which the distribution of the input guides learners in making some inference regarding the number and identity of phonetic categories in the speech stream (Maye and Gerken, 2000; Maye et al., 2002; Kuhl, 2004; Werker et al., 2012). Learners exposed to a bimodal distribution of tokens along some phonetic dimension(s) will infer that there are two phonetic categories each centered at the two peaks of the bimodal distribution, whereas learners exposed to a monomodal distribution will infer that there is only one phonetic category, centered at the peak of the monomodal distribution (see Chapter 2 for further background regarding distributional learning). This section provides an overview of Maye and Gerken (2000), who first provide evidence for distributional learning in an artificial language learning study, since the methodologies of all “A” Experiments are based on this study. Readers who have already read Chapter 2 may wish to skip to Section 2.2.

Maye and Gerken (2000) is the first artificial language learning study which provides experimental support for distributional learning. As detailed in Chapter 2, Maye and Gerken trained adult participants on syllables during an approximately 9-minute training phase. Critical syllables during this training phase began with an alveolar stop taken from an 8-point continuum ranging between a prevoiced stop [d] and a voiceless unaspirated stop [d̥]. Each of these were followed by one of three nuclei, [a æ ə], and had no coda. Participants were either trained on a bimodal distribution or a monomodal distribution of critical syllables. The monomodal distribution of critical syllables consisted of a higher frequency of continuum points near the center of the continuum (points 4 and 5), than those taken from the ends of the continuum (points 1, 2, 7, and 8) (see solid grey line in Figure 15). The bimodal distribution of critical syllables consisted of a higher frequency of continuum points near the endpoints of the continuum (points 2 and 7) than those at the endpoints (points 1 and 8) or in the center (points 4 and 5) (see dotted black line in Figure 15). For simplicity, continuum points will be referred to as D<sub>1</sub>-D<sub>8</sub>.

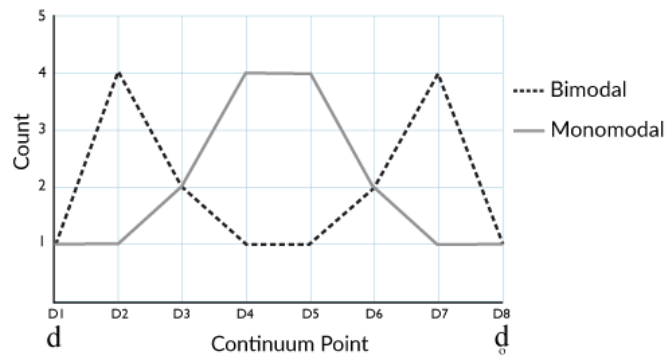


Figure 15. Illustration of the familiarization frequency of onsets of critical stimuli for Bimodal and Monomodal groups during the training phase of Maye and Gerken (2000).

Following training, participants were directed to a test phase in which they were presented with pairs of syllables and were asked if they believed the two syllables presented were the **same** word repeated twice, or two **different** words, in the language they had been exposed to. Maye and Gerken found a greater percentage of “different” responses in the Bimodal condition than in the Monomodal condition when participants were presented with the endpoints of the continuum, D<sub>1</sub> and D<sub>8</sub>. They conclude that this study supports distributional learning.

Maye and Gerken’s findings have led to a number of follow-up experiments by various researchers. Experimental support has been found for adults (Maye and Gerken, 2000; Maye and Gerken, 2001; Hayes-Harb, 2007; Escudero et al., 2011) and infants (by measuring looking times) (Maye et al., 2002; Yoshida, 2010; Maye et al., 2008; Liu and Kager, 2014; Wanrooij et al., 2014; ter Schure et al., 2016). These follow-up experiments on distributional learning have had mixed success. Stimuli successfully used in replications include the stop pairs [d] vs. [ɖ], and [g] vs. [ɡ] (Maye and Gerken, 2000; Maye and Gerken, 2001; Maye et al., 2002; Hayes-Harb, 2007); the vowel pairs [a] vs. [ɑ], and [i] vs. [ɪ] (Gulian et al., 2007; Escudero et al., 2011); and the Thai tone pairs [33] and [241] (Ong et al., 2015). However, Peperkamp et al. (2003) failed to replicate these findings when testing fricatives ranging from [ɸ] to [χ] with French-speaking adult participants. Two out of three experiments presented in Yoshida et al. (2010) (one

with [d] – [ɔ̃] critical onsets, another with [d] – [d̥]) were reported as non-replications with English-learning 10-month old infants. (The reported non-replication in Yoshida et al. (2010) will be discussed further in Section 7.)

Cristia (2018) performed a meta-analysis on distributional learning studies that were conducted on infants. She concludes that there are two types of methodologies that have been utilized in infant distributional learning studies: an alternating/non-alternating design and a habituation/change design.<sup>8</sup> Results of the meta-analysis suggest that infant distributional learning studies which follow an alternating/non-alternating design (e.g. Maye et al., 2002) do not show robust results when taken together, possibly indicating a non-effect or a very small effect size associated with distributional learning. On the other hand, studies which follow a habituation/change design do show a robust effect, supporting the theory that infants make use of distributional learning. Since this dissertation only reports the results of adult studies, I did not utilize either of these methodologies. (Descriptions of each methodology type can be found in the footnote below.)

## 2.2. WEB-BASED EXPERIMENTS

The original purpose of these experiments was simply to determine whether the web is an appropriate platform for conducting distributional learning experiments, as further chapters build on the assumption that distributional learning can be replicated online. Mechanical Turk (“MTurk”) is an online participant

---

<sup>8</sup> Both types of methodologies begin with a training phase in which infants are exposed to auditory stimuli (where critical syllables are presented with either a bimodal or monomodal frequency distribution). In an alternating/non-alternating design, infants are then exposed to alternating or non-alternating trials during the test phase. Alternating trials consist of a string of critical syllables (e.g. continuum points 1 and 8) played one after another in an alternating fashion. Non-alternating trials consist of a string of just one of these continuum points (e.g. either continuum point 1 or continuum point 8), played repeatedly. Experimenters would then analyze the difference in looking times to alternating and non-alternating trials, under the assumption that the greater the difference in looking times to these two types of trials, the more infants are distinguishing between continuum points 1 and 8.

In a habituation/change design, test phases consist of habituating infants to one critical token (e.g. playing continuum point 1 continuously) in a habituation trial. Once infants habituate to the token, a change trial begins and a different critical token is played (e.g. continuum point 8). The experimenter analyzes the difference between looking times in habituation trials and change trials.

pool hosted by Amazon, an electronic commerce company. MTurk began as an online labor market for tasks which are difficult to automate, such as tagging photographs with keywords or selecting the best image out of a sample of images to represent some product. However, an increasing number of researchers have turned to MTurk as a means of recruiting participants (Shapiro et al., 2013; Schnoebelen and Kuperman, 2010). A number of papers have addressed the validity of conducting psychological experiments or surveys online (Denby et al., 2017; Crump et al., 2013; Gosling et al., 2004; Schnoebelen and Kuperman, 2010), and conclude that in-lab and online studies produce similar results (although see Paolacci and Chandler (2014) for warnings regarding the representativeness of the MTurk participant pool to the general population). Denby et al. (2017) find no significant difference in results between a phonotactics learning experiment conducted in-lab and online. Crump et al. (2013) replicate results from a number of common psychological tasks on MTurk. Schnoebelen and Kuperman (2010) compare results from online and in-lab experiments for two linguistics experiments (a word prediction sentence completion task and a semantic similarity judgment task) and also conclude that in-lab and online results were similar.

Although these results are encouraging, the types of tasks reviewed in these studies do not require the ability to detect the fine phonetic detail required in a distributional learning task. Only a few studies have been conducted online which also rely on the ability to distinguish phonetic contrasts. Kleinschmidt (2017) finds that MTurk is suitable for at least some speech perception experiments, in an experiment involving synthetic stimuli in which onsets fall along a continuum between /b/ and /p/, with each of the continuum steps differing in VOT in 10 ms increments (also see Kleinschmidt and Jaeger, 2012; Kleinschmidt and Jaeger, 2015 for further speech perception experiments conducted on Mechanical Turk). The experiments presented in this chapter use critical stimuli drawn from (1) a continuum ranging from voiceless unaspirated [g] (as in *skill* rather than *kill*) to prevoiced [g] (as in *gill*), a continuum which has been successfully implemented in (in-lab) distributional learning experiments such as Maye and Gerken (2001) as well as Hayes-Harb (2007), and (2) a continuum ranging from a voiceless alveopalatal fricative [ç] to a

voiceless retroflex fricative [ʂ], a continuum which has been successfully implemented in an allophony learning experiment, Noguchi (2016).

Since the experimenter cannot be sure that the participant is wearing headphones or even listening during an online experiment, adapting previous distributional learning experiments to an online platform gave rise to a few methodological considerations. To ensure participants were paying attention, catch trials were initially added to both the training and test phases in order. These trials required participants to take some sort of action upon hearing them (e.g. pressing a “1” or a “2” depending on how many tones they had heard). Details of these methodological considerations and their effect on experimental outcomes will be discussed in Section 10.

While this background section has so far provided context regarding the methodological contribution of this chapter, the remaining subsections will focus on the theoretical contribution of this chapter, by first defining two terms: bias and sensitivity.

### 2.3. RESPONSE BIAS VS. SENSITIVITY

This subsection first argues for a distinction to be made between two phenomena: **response bias** (also simply “bias”) and **sensitivity**. Following this, analysis methods used by previous experimental studies on distributional learning are reviewed to highlight the fact that past studies differ in which metrics are used as supporting evidence for distributional learning, and that the metrics used measure different phenomena.

To illustrate the difference between response bias and sensitivity, suppose a subject is given the task of determining whether two marbles drawn randomly from a bag are the same color or are two different colors. Half of the participants are told that there exist two shades of green in the bag while the other half of the participants are not told anything about the number of shades of green. In reality though, there is only one shade of green marble. In this scenario, one might imagine that the participants who are told that there are two different shades of green in the bag are more likely to respond that two green marbles are different shades compared to participants who are not told anything about the number of shades of

green in the bag, even though all participants are seeing the same exact same shade of green. In this first scenario, we can say that the group of participants told that there are two shades of green is more **biased** towards a “different” response compared to the group of participants who had not been told anything regarding the number of shades of green in the bag. Even when faced with two marbles with the exact same shade, participants with a bias towards a “different” response are more likely to respond “different” compared to the participants who are not told anything about the number of shades of green, simply because they are expecting there to be two shades of green in the bag.

Although this is a hypothetical example, factors which affect response bias have been identified (Macmillan and Creelman, 2004; Stretch and Wixted, 1998; Dougal and Rotello, 2007; Baddeley and Colquhoun, 1969; See et al., 1997; Davenport, 1969). For example, Dougal and Rotello (2007) presented participants with a list of words during a training phase. In a following test phase, participants were presented with words that had been presented during testing, as well as words which had not been presented during testing. Participants were asked to determine whether they encountered each word in the training phase. Dougal and Rotello found that, regardless of whether or not a word was actually encountered during training, participants were more likely to respond that they recalled the word from the training phase if the word was a negative emotion-related word, compared to positive or neutral words. Therefore participants had a greater bias towards responding that a word was encountered before if it was a negative emotion-related word than if it was a positive or neutral word. Similarly, in a task involving detection of a vibration through a transmitter placed around participants’ arms, participants’ response biases were found to be affected by a scoring system based on their performance. If more points were subtracted for missing a detection (Signal Detection Theory terminology: “miss”) or responding when there was no signal (SDT: “false alarm”), participants were more conservative in their responses, exhibiting lower response biases towards responding that there had been a vibration, but the same sensitivity ( $d'$ ) to stimuli (Davenport, 1969). Response bias also changes based on the probability that some signal is encountered (See et al.,

1997). Participants who encountered some signal at a low probability were more conservative in their responses compared to participants encountering the signal at a high probability (Baddeley and Colquhoun, 1969; See et al., 1997).

Returning to the marble example, we can imagine another scenario, identical to the one above, except this time two different shades of green actually exist in the bag. In this scenario, one might imagine that the group that had been told that two different shades of green exist has become better at detecting whether two green marbles are the same or different shades compared to the other group. That is, their **sensitivity** to the slight difference in shade is heightened because they know a difference exists, and their heightened awareness makes them more sensitive of the different shades. In this scenario, the group with the higher sensitivity responds “different” more often when the two marbles are actually different shades of green, but crucially also responds “same” more often when the two marbles are the same shade of green. In the previous scenario, the group with the lower response bias responded “different” more often even when the two marbles were the same shade of green. Sensitivity differences have been found in numerous tasks. For example, trained radiologists show higher sensitivity to low-contrast dots on X-rays than novice students (Sowden et al., 2000). Iverson and Kuhl (1995) find that participants have lower sensitivities (measured in  $d'$ ) to pairs of stimuli near prototypical vowel centers than those near non-prototypical vowels. Guion et al. (2000) find that native English speakers have significantly greater sensitivity to the contrast between [ɹ] and [l] compared to native Japanese speakers, while native Japanese speakers have greater sensitivity to the contrast between [r] and [d] (measured in Signal Detection Theory's  $A'$ , an estimate of the area under an ROC curve which does not require the assumption that the response variable is normally-distributed (Stanislaw and Todorov, 1999). They also find that native Japanese speakers with more English experience have higher sensitivities compared to Japanese speakers with less English experience to some contrasts made in English, like /ɪ/ and /w/, but not to other contrasts, like /ɪ/ and /l/. A number of distributional learning studies find that groups trained on a bimodal distribution

have a higher sensitivity (measured in  $d'$ ) to critical stimuli after training than a monomodal or control group (Hayes-Harb, 2007; Noguchi, 2016).

The scenario of drawing two marbles from a bag is similar to the scenario presented to a participant in a distributional learning experiment. During the test phase, the participant is asked to determine whether two sounds, rather than marbles, are the same or different. One can imagine two different theories regarding the underlying mechanism behind distributional learning, each of which makes different predictions regarding sensitivity and response bias, stated below.

- 1) **The Sensitivity Hypothesis:** Distributional learning causes **sensitivity** to improve, such that phones which are presented in a bimodal distribution to the learner are perceived as being more distinct from one another than phones which are presented in a monomodal distribution. The more perceptually distinct two given sounds are, the more likely a learner is to believe that these two sounds are “different.” Therefore, a bimodal distribution results in increased discriminatory ability (sensitivity) between the endpoints. Note that a change in response bias may occur later. A schematic of this is shown in Figure 16.
- 2) **The Bias Hypothesis:** Bimodal training results in learners being more likely to think that two sound categories exist in the speech stream compared to monomodal training. Therefore, the bimodal learners have a greater **bias** towards responding “different.” For example, a learner is more likely to think there must be two “g”-like sounds in the speech stream if exposed to a bimodal distribution of “g” sounds. Perception, however, is not directly affected by distributional learning (although a change in sensitivity may occur later). A schematic of this is shown in Figure 17.



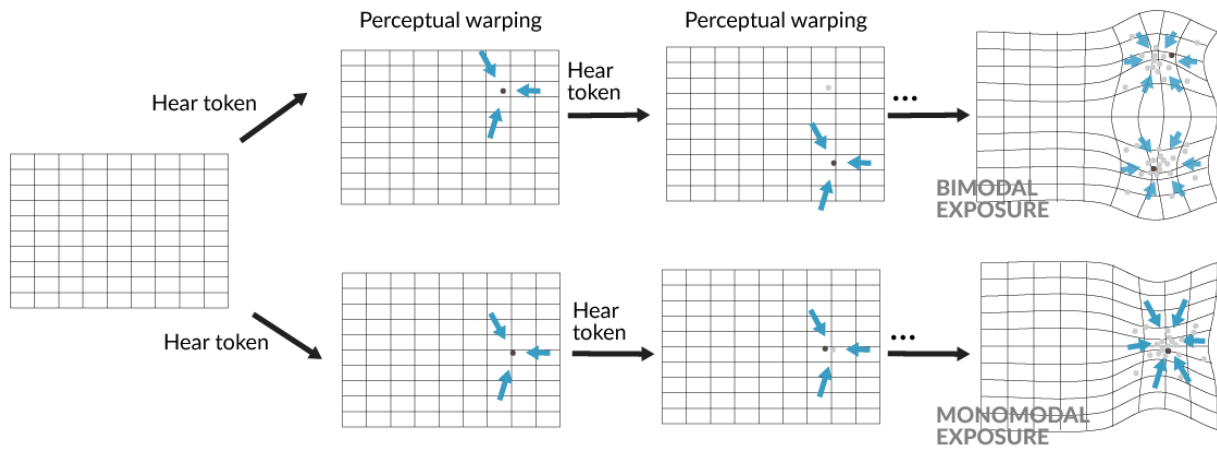


Figure 16. Illustration of the Sensitivity Hypothesis.

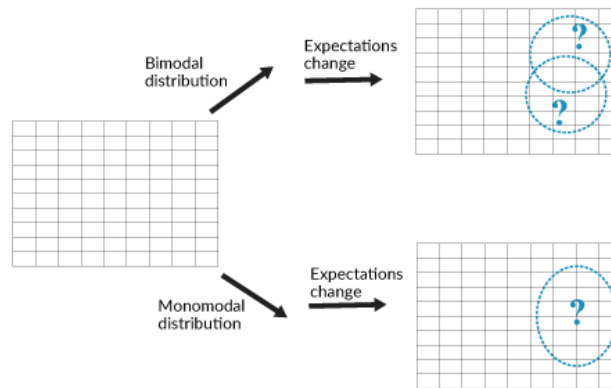


Figure 17. Illustration of the Response Bias Hypothesis.

When sensitivity increases, it is expected that learners will be better at determining that two different stimuli are in fact different, as well as better at determining that two identical stimuli are the same. However, when the bias to respond “different” increases, it is expected that learners will respond that pairs of stimuli are different more often, regardless of whether they are in fact the same or different. This is summarized in Table 1.

	Percentage of “different” responses for Different Pairs (e.g. D <sub>1a</sub> vs. D <sub>8a</sub> )	Percentage of “different” responses for Same Pairs (e.g. D <sub>1a</sub> vs. D <sub>1a</sub> )
Increased bias towards “different” response	↑	↑
Increased sensitivity	↑	↓
Increased bias toward “different” response and increased sensitivity	↑↑	↑

Table 1. Effect of increased bias towards a “different” response compared to effect of increased sensitivity on a learner's responses. The third scenario, in which both response bias and sensitivity are affected, is shown here, but not discussed, as no evidence is found for it in the set of experiments conducted in this chapter.

Previous metrics used to measure distributional learning either measure sensitivity, or are inconclusive in whether they measure sensitivity or response bias. Maye and Gerken (2000) analyze the percentage of times participants respond that critical Different Pairs are “different,” finding that the participants trained on a monomodal distribution respond “different” less often than participants trained on a bimodal distribution. However, as can be seen in Table 1, this finding is compatible with an increase in sensitivity *or* in an increase in bias towards a “different” response. Since Maye and Gerken only analyze the effect on Different Pairs, it is unclear whether sensitivity or response bias is the factor being affected. Noguchi (2016) on the other hand measures sensitivity in the form of d-prime values. He finds that participants in the control condition have lower d-prime values than participants trained on a bimodal distribution. The current study will analyze both sensitivity and response bias. (Noguchi (2016) does not report on response bias.)

The distinction between sensitivity and response bias is an important one to make, because perceptual warping accounts of distributional learning (such as those suggested by Guenther and Gjaja (1996) and Boersma et al. (2003)) predict that sensitivity will *always* be affected by distributional learning, since a change in sensitivity is the underlying cause of distributional learning in these models. The current study measures both sensitivity and response bias, and will conclude that both can be affected by distributional learning. I argue that direct perceptual warping accounts of distributional learning are not

supported by this study. Rather, a change in response bias occurs at an initial stage of distributional learning, followed by a change in sensitivity at a second stage.

#### 2.4. DISTRIBUTIONAL LEARNING IN FUNCTIONAL PHONOLOGY

In Chapter 2, two models of distributional learning were described: one by Boersma et al. (2003) and the other by Guenther and Gjaja (1996). Both of these models predict that perceptual warping always accompanies distributional learning. This section will briefly summarize the main ideas in Boersma et al.'s Optimality Theoretic model of distributional learning, and then show how these concepts translate to the *same-different* task in Maye and Gerken (2000).

Boersma et al. (2003) models distributional learning within Functional Phonology, a framework which attempts to explain theories of phonology with functional phonetic principles, such as minimization of perceptual confusion (Boersma, 1998). In order to see how the model in Boersma et al. (2003) can be extended to the task in Maye and Gerken (2000), suppose the learner is faced with a distribution of 8 stops ranging in prevoicing from 0 ms to 140 ms in 20 ms intervals.

According to Boersma et al.'s model, training exposure to a bimodal distribution of these stops will lead the learner to be more sensitive to the endpoints of the continuum, as opposed to training exposure to a monomodal distribution. In terms of a *same-different* task, this would mean that bimodal training results in more “different” responses to Different Pairs, and fewer “different” responses to Same Pairs than monomodal training. In Boersma et al.'s model, there are two levels of representation: the acoustic representation, which makes up the input; and the phonetic categorization, which makes up possible candidate outputs. There are three families of constraints responsible for distributional learning: \*CATEG, PERCEIVE, and \*WARP. The \*CATEG(ORIZE) constraints punish perceptual categories with some particular acoustic value. For example, a high-ranked \*CATEG (/0 ms/) would have the effect of prohibiting categorization of some input into a category with a prevoicing value of /0 ms/. PERCEIVE constraints require the listener to perceive (categorize) an auditory input with a particular acoustic value as a member of *some* category, so the null candidate /-/ violates PERCEIVE ([0 ms]) if the input was [0 ms]. \*WARP

constraints are violated if the difference between the acoustic value of the input and that of the output is greater than the amount defined by the \*WARP constraint. For example, \*WARP (20 ms) is violated if the difference between the acoustic value of the input (e.g. [0 ms]) and the acoustic value of the candidate (e.g. /20 ms/) is equal to or greater than 20 ms. Initially, all \*CATEG constraints are ranked high, and all PERCEIVE constraints are ranked low.

[20 ms]	*CATEG (/0 ms/)	*CATEG (/20 ms/)	PERCEIVE ([20 ms])	*WARP (20 ms)
/0 ms/	*!			*
/20 ms/		*!		
☑ /-/			*	

Figure 18. Tableau illustrating the initial state in which \*CATEG constraints are ranked high and PERCEIVE constraints are ranked low.

As seen in Figure 18, initially an input value of [20 ms] would be perceived as the null candidate /-/ , due to the higher-ranked \*CATEG constraints. According to this model, the learner is unsatisfied with perceiving the null category, and so will categorize the input value as its identical counterpart value, in this case, /20 ms/. This target winner is indicated in the tableaux with a check mark. Knowing this target, the learner now reranks their constraints, in this case lowering the \*CATEG (/20/) constraint and raising the PERCEIVE ([20]) constraint (see Figure 19). Over time, PERCEIVE ([20]) will outrank \*CATEG (/20/).

[20 ms]	*CATEG (/0 ms/)	*CATEG (/20 ms/)	PERCEIVE ([20 ms])	*WARP (20 ms)
/20 ms /				
/0 ms/	*!			*
✓ /20 ms /		*! →		
☑ /-/			← *	

Figure 19. An example of \*CATEG demotion and PERCEIVE promotion. Figure from Boersma et al. (2003).

Following the Gradual Learning Algorithm, reranking occurs in small increments. After hearing many [20 ms] values, the learner's \*CATEG (/20 ms/) constraint will be lower than the learner's PERCEIVE ([20 ms]), resulting in the winning candidate /20 ms/. Therefore at this point, the learner categorizes an input of [20 ms] as belonging to a /20 ms/ category.

If learners heard all 8 prevoicing values in identical relative amounts, the learner would simply perceive all incoming values as belonging to its own category. However, since the learner does *not* receive a non-modal input of prevoicing categories, various \*CATEG constraints will outrank other \*CATEG constraints. For example, if a learner hears more stops with [20 ms] of prevoicing than stops with [0 ms] prevoicing, then the PERCEIVE and \*CATEG constraints associated with a 20 ms prevoiced token (PERCEIVE [20] and \*CATEG /20/) will move *more* than the PERCEIVE and \*CATEG constraints associated with the stop with 0 ms prevoicing token (PERCEIVE [0] and \*CATEG /0/). Crucially, this will lead to \*CATEG constraints associated with more commonly-heard tokens to be ranked beneath \*CATEG constraints associated with less commonly-heard tokens. Therefore in this model, the nonuniform nature of the input distribution results directly in perceptual warping.

[0 ms] /0 ms/	PERCEIVE ([20 ms])	PERCEIVE ([0 ms])	*CATEG (/0 ms/)	*CATEG (/20 ms/)	*WARP (20)
√ /0 ms/			*! →		
⊗ /20 ms/				← *	← *
/-/		*!			

Figure 20. Sample tableau for listener who has heard more [20 ms] tokens than [0 ms] tokens. Upon hearing a [0 ms] token, this listener will perceive it as being /20 ms/.

So far this section has described the basic framework behind the Functional Phonology model of distributional learning. The following subsections will attempt to extend this model so that we can see what predictions it makes when faced with a *same-different* task. To preview, Sections 2.4.1-2.4.3 can be summarized as follows: the theoretical end state of the learner as described by Boersma et al. (2003) predicts increased sensitivity to stimuli in a *same-different* task (Section 2.4.1). However, whether the learner can arrive at this theorized end state is another question. Section 2.4.2 provides evidence suggesting that modelling the learning process with the Gradual Learning Algorithm and by choosing parameters carefully seems capable of resulting in the theorized end state used in Section 2.4.1. However, as noted in Section 2.4.3, the choice of these parameters appears to be arbitrary, and this learner does not naturally converge on the end state grammar used in Section 2.4.1.

#### 2.4.1 Boersma et al. (2003): From (Theoretical) End State Rankings to Same-Different Experiment

This section translates the theoretical end state constraint rankings of Boersma et al. (2003)'s model to a *same-different* task in order to determine what predictions this model of distributional learning makes in the task given in Maye and Gerken (2000). It will be shown that there exists a choice of end state constraint rankings which predicts an increase in sensitivity.

Recall that Maye and Gerken (2000)'s participants were presented either with Same Pairs, which were identical pairs of syllables taken from the endpoints of an 8-point continuum, or with Different Pairs, which were syllables taken from the opposite ends of the 8-point continuum. The A1 and A2 Experiments described later will make use of a velar stop continuum ranging between [g̊], which has 0 ms of prevoicing, and [g], which has 140 ms of prevoicing. Participants will be asked to respond "same" if they believe the pair consists of the same syllables, and "different" if they believe the pair consists of two different syllables. Therefore Same Pairs consist of either a [0 ms]/[0 ms] pair of syllables, or a [140 ms]/[140 ms] pair. Different Pairs consist of either a [0 ms]/[140 ms] pair of syllables, or a [140 ms]/[0 ms] pair. A participant will respond "same" in this *same-different* task if they map both inputs of the trial pair to the same output; otherwise they will respond "different."

Although exact guidelines for how initial constraint rankings should be determined for a given stimulus set are not given, Boersma and colleagues provide a set of initial constraint rankings for a hypothetical example involving vowel heights. This section will follow the general range of numbers that Boersma and colleagues use in their initial state, but, as they do not provide rankings after learning simulations have occurred, this section speculates on the constraint rankings in the final state. It will be shown that constraint rankings can be chosen so that a bimodally-trained learner has greater sensitivity to the endpoints of the stimulus set compared to a monomodally-trained learner, as determined in a *same-different* task.

Boersma and colleagues arrange the initial ranking of the \*WARP family of constraints so that the greater the amount of warping, the higher the constraint’s initial rank. For example, \*WARP(100) is initially ranked at 100, and \*WARP(120) is initially ranked at 120. This reflects a greater penalty for perceiving an output the more it veers from its original input and can be accomplished by fixing rankings of higher \*WARP constraints above lower \*WARP constraints. Boersma and colleagues also distinguish between very low-warping \*WARP constraints, and rank all of these at the very low value of  $-10^9$ . They justify this in terms of psychoacoustics: some differences fall below the Just-Noticeable Difference value (JND), and so are not distinguishable by participants. This group of initially low-ranked constraints will be referred to as “bottom-ranked” \*WARP constraints.

The level of warping allowed in this model will determine how many Different Pairs are mistakenly categorized as being the “same” by a participant. If all \*WARP constraints which warp the input by 60 or fewer ms are categorized as “bottom-ranked” \*WARP constraints, and all other \*WARP constraints are high-ranked, an input of [0 ms] will not be categorized as any input which differs by more than 60 ms. Therefore [0 ms] can have a possible output of /0/, /20/, /40/, and /60/, but not of /80/ since \*WARP(80) is a high-ranked constraint. Similarly, for an input of [140 ms], possible outputs include /140/, /120/, /100/, and /80/, but not /60/ (see Figure 21, left).

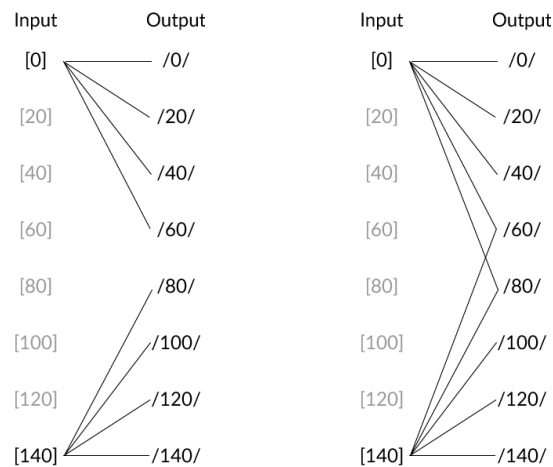


Figure 21. (left) Possible outputs for an input of [0] and [140] when all \*WARP constraints including and below \*WARP(60) are very low ranked. (right) Possible outputs for an input of [0] and [140] when all \*WARP constraints including and below \*WARP(80) are very low ranked.

If, however, \*WARP(80) is included as one of the “bottom-ranked \*WARP constraints,” an input of [0 ms] is allowed to have an output of /80/, and an input of [140 ms] is allowed to have an output of /60/ (see Figure 21, right).

If we wish to allow *some* incorrect “same” responses to a Different Pair, then we will need an input of [140] and an input of [0] to occasionally have the same output. This is only possible if we allow \*WARP(80) to be included in the set of “bottom-ranked” \*WARP constraints. For this simulation, \*WARP(80) will be our cut-off point, and all \*WARP constraints above this will be high-ranked constraints. This section uses the constraint rankings shown in Table 2 to simulate the end state for a bimodally-trained learner (left) and monomodally-trained learner (right).

<b>Constraint</b>	<b>Ranking (Bimodal)</b>	<b>Constraint</b>	<b>Ranking (Monomodal)</b>
*WARP (/140 ms/)	140	*WARP (/140 ms/)	140
*WARP (/120 ms/)	120	*WARP (/120 ms/)	120
*WARP (/100 ms/)	100	*WARP (/100 ms/)	100
PERCEIVE (/20 ms/)	40	PERCEIVE (/60 ms/)	40
PERCEIVE (/120 ms/)	40	PERCEIVE (/80 ms/)	40
PERCEIVE (/40 ms/)	20	PERCEIVE (/40 ms/)	20
PERCEIVE (/100 ms/)	20	PERCEIVE (/100 ms/)	20
PERCEIVE (/0 ms/)	10	PERCEIVE (/0 ms/)	10
PERCEIVE (/60 ms/)	10	PERCEIVE (/20 ms/)	10
PERCEIVE (/80 ms/)	10	PERCEIVE (/120 ms/)	10
PERCEIVE (/140 ms/)	10	PERCEIVE (/140 ms/)	10
*CATEG (/140 ms/)	-10	*CATEG (/140 ms/)	-10
*CATEG (/80 ms/)	-10	*CATEG (/120 ms/)	-10
*CATEG (/60 ms/)	-10	*CATEG (/20 ms/)	-10
*CATEG (/0 ms/)	-10	*CATEG (/0 ms/)	-10
*CATEG (/100 ms/)	-20	*CATEG (/100 ms/)	-20
*CATEG (/40 ms/)	-20	*CATEG (/40 ms/)	-20
*CATEG (/120 ms/)	-40	*CATEG (/80 ms/)	-40
*CATEG (/20 ms/)	-40	*CATEG (/60 ms/)	-40
*WARP (80 ms)	-10 <sup>9</sup>	*WARP (80 ms)	-10 <sup>9</sup>
*WARP (60 ms)	-10 <sup>9</sup>	*WARP (60 ms)	-10 <sup>9</sup>
*WARP (40 ms)	-10 <sup>9</sup>	*WARP (40 ms)	-10 <sup>9</sup>
*WARP (20 ms)	-10 <sup>9</sup>	*WARP (20 ms)	-10 <sup>9</sup>

Table 2. Possible end state ranking for bimodally-trained learner (left) and monomodally-trained learner (right).



Note that all \*WARP constraints have the same ranking for both learners. PERCEIVE and \*CATEG constraints associated with most frequently-heard tokens ([20] and [120] for the bimodally-trained participant; [60] and [80] for the monomodally-trained participant) have shifted the most during training, and so are the furthest apart from one another. To take the bimodal learner for example, PERCEIVE(/20/) and \*CATEG([20]) are both associated with the frequently-heard [20]. Because of this, they have moved more than the constraints PERCEIVE(/140/) and \*CATEG([140]). Specifically, PERCEIVE(/20/) and \*CATEG([20]) are ranked at 40 and -40 respectively, which is more distant than the relative ranking of PERCEIVE(/140/) and \*CATEG([140]), which are ranked at 10 and -10.

In order to test the predictions that the Boersma et al. model makes on an artificial language learning experiment modelled on Maye and Gerken (2000), a given set of constraint rankings will be evaluated by using the OT Learning tools available in Praat 6.0.29 (Boersma, 2002), speech analysis software which also has functions for evaluating Optimality Theoretic grammars and running learning simulations. Praat was provided with the constraint rankings shown in Table 2, and was asked to simulate input-output pairs, with inputs chosen randomly (and uniformly) from the 8 possible inputs [0 ms], [20 ms]...[140 ms]. Outputs were evaluated with an evaluation noise of 40. The non-zero evaluation noise assumes that each constraint is represented by a Gaussian probability distribution centered around final ranking values with a standard deviation of 40.

To translate these input-output pairs to a *same-different* task, this section calculates the probability that a given input is categorized as each of the 8 possible prevoicing categories (/0 ms/, /20 ms/, ... /140 ms/), for the two inputs corresponding to the two endpoints of the continuum, [0 ms] and [140 ms]. This would correspond to the probability of categorizing input as /G<sub>n</sub>/ given an input of [G<sub>1</sub>], or  $P(/G_n/ | [G_1])$ , and the probability of categorizing input as /G<sub>n</sub>/ given an input of [G<sub>8</sub>], or  $P(/G_n/ | [G_8])$ . In order to calculate the probability of responding “same” or “different” on a *same-different* task, the probability of an output of each prevoicing category given an input of [0 ms] or [140 ms] would need to be calculated. These are shown in Figure 22.

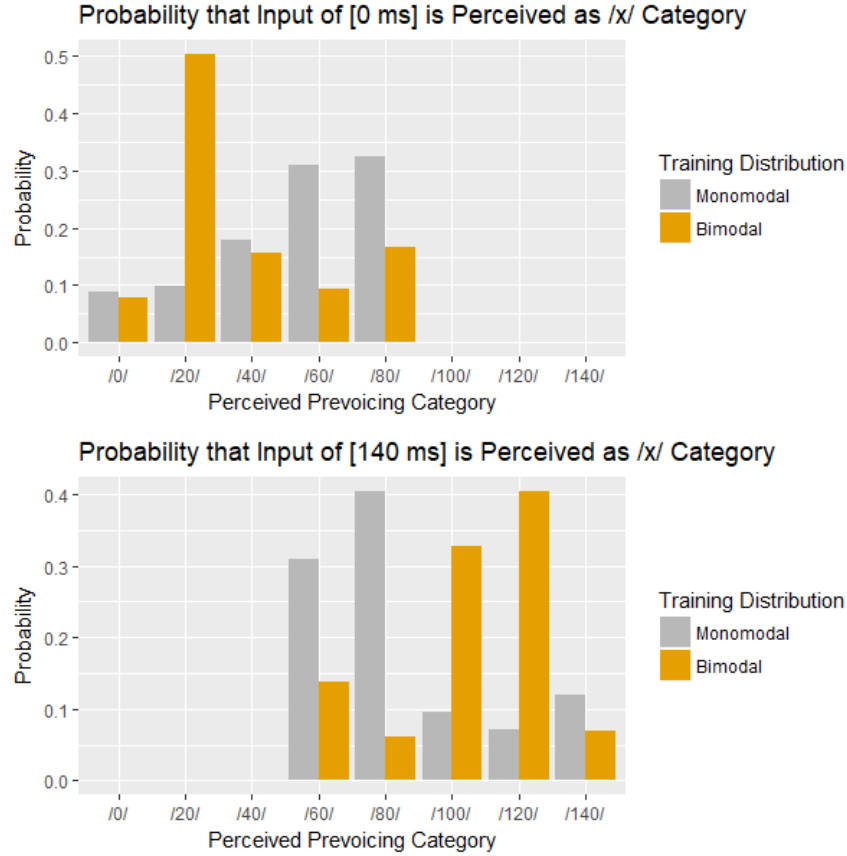


Figure 22. Probability that an input of [0 ms] (top) and an input of [140 ms] (bottom) are categorized in each of the 8 possible prevoicing categories.

For the purposes of this chapter, this section is only interested in the *comparison* between monomodally- and bimodally-trained grammars, rather than the exact probability values. The probability of a “different” response on a Different Pair is calculated using Equation (1); the probability of a “different” response on a Same Pair is calculated using Equation (2).

$$(1) \text{ DiffPair prob. of "d" resp.} = 1 - \sum_{n=1}^8 P(/G_n/ | [G_1]) * P(/G_n/ | [G_8])$$

$$(2) \text{ SamePair prob. of "d" resp.} = 1 - \frac{1}{2} * \left( \sum_{n=1}^8 P(/G_n/ | [G_1]) * P(/G_n/ | [G_1]) + \sum_{i=1}^8 P(/G_n/ | [G_8]) * P(/G_n/ | [G_8]) \right)$$

In Equation (1), the  $P(/G_n/ | [G_1]) * P(/G_n/ | [G_8])$  component calculates the probability that both inputs  $[G_1]$  and  $[G_8]$  were categorized as the same category  $/G_n/$ . This is done for all 8 possible categories and

summed up. This summation represents the probability of a “same” response if a participant is presented with a Different Pair ( $G_1$  vs.  $G_8$ ). This is subtracted from 1 to represent the probability of a “different” response. The equation in (2) follows similar logic, and simply calculates the average of two [ $G_1$ ] inputs being categorized as the same category with the probability of two [ $G_8$ ] inputs being categorized as the same category. The percentage of “different” responses for Same Pairs and Different Pairs are presented in Table 3.

	Same Pair	Diff Pair
Monomodal	73.1%	77.3%
Bimodal	69.0%	97.7%

Table 3. Predicted percentage of “different” responses in a *same-different* task based on the rankings given in Table 2.

Exact percentages are not especially important here. These could be altered to fit a given study’s results by changing parameters of the evaluation (such as the evaluation noise) or by changing the relative distance between constraint rankings. What should be noted here is the differences in percentages between the monomodally- and bimodally-trained grammars. The bimodally-trained grammar responds that fewer Same Pairs are “different” and more Different Pairs are “different” compared to the monomodally-trained grammar. In other words, the bimodal group is more sensitive than the monomodal group to stimuli and therefore to the difference between Same Pairs and Different Pairs. Therefore, according to this model, nonuniform frequency distribution directly leads to perceptual warping, such that the bimodal group has a greater *sensitivity* to stimuli than the monomodal group.

#### 2.4.2 Boersma et al. (2003): From Initial State to End State

In order to test whether the constraint rankings found in the previous section can be learned, this section will simulate this model with the Gradual Learning Algorithm as implemented in Praat 6.0.29 (Boersma, 2002). The Gradual Learning Algorithm is an error-driven model for constraint-based grammars which updates rankings of constraints in small (“gradual”) increments. At each iteration, the GLA selects an input and provides an output based on its current ranking. “Learning” occurs when there is a mismatch

between this output and the target output. Using the above example of a learner faced with a distribution of 8 stops ranging in prevoicing from 0 ms to 140 ms in 20 ms intervals, the learner’s initial state consists of 23 constraints, given in Table 4. Each constraint is given an initial ranking and a plasticity value. Initial rankings determine how high or low a constraint starts out at the beginning of learning. Boersma et al. (2003) state that all \*CATEG constraints initially outrank all PERCEIVE constraints. Therefore all \*CATEG constraints were given an initial ranking of 1, and all PERCEIVE constraints were given an initial ranking of -1. “Bottom-ranked” \*WARP constraints were given the initial ranking of  $-10^9$ , while all other \*WARP constraints were given the initial rankings of 100 or more (\*WARP(100) at 100; \*WARP(120) at 120; \*WARP(140) at 140). Plasticity defines how much or how little a constraint can be promoted or demoted at each iteration of the GLA, with a value of 0 corresponding to a constraint which cannot be promoted or demoted. All constraints were given a plasticity of 1, with the exception of \*WARP constraints, which were given a plasticity of 0. The initial grammar is given in Table 4.

<b>Constraint</b>	<b>Initial Ranking</b>	<b>Plasticity</b>
*WARP (140 ms)	140	0
*WARP (120 ms)	120	0
*WARP (100 ms)	100	0
*CATEG (/140 ms/)	1	1
*CATEG (/120 ms/)	1	1
*CATEG (/100 ms/)	1	1
*CATEG (/80 ms/)	1	1
*CATEG (/60 ms/)	1	1
*CATEG (/40 ms/)	1	1
*CATEG (/20 ms/)	1	1
*CATEG (/0 ms/)	1	1
PERCEIVE (/140 ms/)	-1	1
PERCEIVE (/120 ms/)	-1	1
PERCEIVE (/100 ms/)	-1	1
PERCEIVE (/80 ms/)	-1	1
PERCEIVE (/60 ms/)	-1	1
PERCEIVE (/40 ms/)	-1	1
PERCEIVE (/20 ms/)	-1	1
PERCEIVE (/0 ms/)	-1	1
*WARP (80 ms)	$-10^9$	0
*WARP (60 ms)	$-10^9$	0
*WARP (40 ms)	$-10^9$	0
*WARP (20 ms)	$-10^9$	0

Table 4. Initial state of the learner.

In addition to an initial grammar, the GLA as implemented in Praat also takes an input file. The input file provides the learner with relative frequencies of input tokens, as well as a target winner for each input. At each iteration of the algorithm, the GLA updates its constraint rankings if the actual winner and the target winner do not match.

To obtain simulated end state grammars, the algorithm was initially run twice: once with a monomodal input, and once with a bimodal input. The distribution frequencies matched those used in Maye and Gerken (2000) (see Figure 15). That is, the monomodal input presented the algorithm with the following relative frequencies: [0 ms] =  $n$ , [20 ms] =  $n$ , [40 ms] =  $2n$ , [60 ms] =  $4n$ , [80 ms] =  $4n$ , [100 ms] =  $2n$ , [120 ms] =  $n$ , [140 ms] =  $n$ . The bimodal input presented the algorithm with the following relative frequencies: [0 ms] =  $n$ , [20 ms] =  $4n$ , [40 ms] =  $2n$ , [60 ms] =  $n$ , [80 ms] =  $n$ , [100 ms] =  $2n$ , [120 ms] =  $4n$ , [140 ms] =  $n$ . The GLA chose inputs randomly from these distributions at each iteration.

The GLA was run with the following parameters: Evaluation noise = 2, Update rule = Symmetric all, Initial plasticity = 20, Replications per plasticity = 1,000,000, Plasticity decrement = 0.1, Number of plasticities = 10, Relative plasticity spreading = 0.1, Honour local rankings, Number of chews = 1. The resulting rankings after the GLA was run on a bimodal and monomodal distribution are shown in Table 5.

<b>Final Ranking - Monomodal Training</b>	
<b>Constraint</b>	<b>Ranking</b>
*WARP (140 ms)	140
*WARP (120 ms)	120
*WARP (100 ms)	100
PERCEIVE ([20 ms])	20.7
PERCEIVE ([120 ms])	20.0
PERCEIVE ([0 ms])	19.8
PERCEIVE ([140 ms])	18.5
PERCEIVE ([80 ms])	-1.0
PERCEIVE ([40 ms])	-1.0
PERCEIVE ([60 ms])	-1.0
PERCEIVE ([100 ms])	-1.0
*CATEG (/20 ms/)	-8.8
*CATEG (/120 ms/)	-8.8
*CATEG (/140 ms/)	-8.9
*CATEG (/0 ms/)	-8.9
*CATEG (/100 ms/)	-9.5
*CATEG (/40 ms/)	-9.5
*CATEG (/60 ms/)	-10.3
*CATEG (/80 ms/)	-10.3
*WARP (80 ms)	-10 <sup>9</sup>
*WARP (60 ms)	-10 <sup>9</sup>
*WARP (40 ms)	-10 <sup>9</sup>
*WARP (20 ms)	-10 <sup>9</sup>

<b>Final Ranking - Bimodal Training</b>	
<b>Constraint</b>	<b>Ranking</b>
*WARP (140 ms)	140
*WARP (120 ms)	120
*WARP (100 ms)	100
PERCEIVE ([0 ms])	23.0
PERCEIVE ([140 ms])	18.3
PERCEIVE ([20 ms])	17.9
PERCEIVE ([120 ms])	16.9
PERCEIVE ([100 ms])	-1.0
PERCEIVE ([40 ms])	-1.0
PERCEIVE ([60 ms])	-1.0
PERCEIVE ([80 ms])	-1.0
*CATEG (/60 ms/)	-8.0
*CATEG (/80 ms/)	-8.0
*CATEG (/0 ms/)	-8.7
*CATEG (/140 ms/)	-8.7
*CATEG (/100 ms/)	-8.9
*CATEG (/40 ms/)	-8.9
*CATEG (/20 ms/)	-10.4
*CATEG (/120 ms/)	-10.4
*WARP (80 ms)	-10 <sup>9</sup>
*WARP (60 ms)	-10 <sup>9</sup>
*WARP (40 ms)	-10 <sup>9</sup>
*WARP (20 ms)	-10 <sup>9</sup>

Table 5. End state constraint rankings after monomodal training (left) and bimodal training (right).

After training, all PERCEIVE constraints outrank \*CATEG constraints, allowing the learner to categorize input. The crucial rankings to note here are those among the \*CATEG family. Recall that the peak of the monomodal distribution is located at [60 ms] and [80 ms], and that the peaks of the bimodal distribution are located at [20 ms] and [120 ms]. Note the \*CATEG constraints associated with those prevoicing categories (highlighted in grey in Table 5). After monomodal training, the \*CATEG constraints of VOT values corresponding to peaks of the bimodal distribution, \*CATEG (/120 ms/) and \*CATEG (/20 ms/), are higher-ranked than the \*CATEG constraints of VOT values corresponding to the peak of the monomodal distribution, \*CATEG (/60 ms/) and \*CATEG (/80 ms/). However, after bimodal input, the \*CATEG constraints of VOT values corresponding to peaks of the bimodal distribution are *lower*-ranked than the \*CATEG constraints of VOT values corresponding to the peak of the monomodal distribution. In other words,

constraints corresponding to tokens presented to learners more often were promoted more during constraint re-ranking, resulting in lower-ranked corresponding \*CATEG constraints. This simulation was run three more times with the same parameter settings to ensure qualitatively similar outcomes. In all cases, \*CATEG(/20 ms/) and \*CATEG(/120 ms/) were higher-ranked than \*CATEG(/60 ms/) and \*CATEG(/80 ms/) in the end state that was trained on a monomodal input, and lower-ranked than \*CATEG(/60 ms/) and \*CATEG(/80 ms/) in the end state that was trained on a bimodal input.

The final rankings given in Table 5 were translated to a *same-different* task, following the method given in the previous subsection. The predicted percentage of “different” responses is given in Table 6.

	Same Pair	Diff Pair
Monomodal	50.82%	99.56%
Bimodal	49.76%	99.64%

Table 6. Predicted percentage of “different” responses in a *same-different* task based on the rankings given in Table 5.

The predicted percentage of “different” responses in a *same-different* task for the learned end state constraint rankings show only a slight difference between monomodal and bimodal training, especially when compared to the predicted responses given in Table 3, which was derived from contrived end state constraint rankings (specifically, those given in Table 2). However, we see that the difference in percentages trend in the same direction as those shown in Table 3: the bimodally-trained end state is better at determining that Same Pairs are the same than the monomodally-trained end state (as shown by the slightly lower percentage of “different” responses to Same Pairs), but better at determining that Different Pairs are different (as shown by the slightly higher percentage of “different” responses to Different Pairs). Therefore, although I was not able to come up with a set of parameters which resulted in *same-different* response percentages that are as distinct as those derived from the contrived end state constraint rankings, it does appear that a set of parameters exists which would result in similar *same-different* responses as those shown in Table 3.

### 2.4.3 Boersma et al. (2003): Weaknesses of the Model

To summarize, Section 2.4.1 showed that 1) the constraint rankings proposed in Boersma et al. (2003) can be translated into the results of a *same-different* experiment similar to that used by Maye and Gerken (2000), and 2) there exists an end state of constraint rankings which results in greater sensitivity to the endpoint stimuli as determined in a *same-different* experiment. Specifically, the constraints proposed in Boersma et al. (2003) can be ranked in a way so that sensitivity as observed in a *same-different* task is greater for a bimodally-trained learner compared to a monomodally-trained learner. Section 2.4.2 attempted to show that the end state could be arrived at with the correct choice of beginning state constraint rankings and parameters. It showed that the general idea of the theoretical end state used in Section 2.4.1 could be arrived at, but did not develop a principled way of setting these initial rankings and model parameters.

The greatest weakness is that the model put forth in Boersma et al. (2003) does not put forward a set of principles to determine initial rankings and parameters. This is not a trivial matter, as parameters relating to plasticity and amount of training exposure the model was given would have a great effect on the end state grammar. Additionally, attempts at simulations which are not reported here appear to indicate that evaluation noise and the distances between constraint rankings seem to have a large impact on the exact percentages of predicted “different” responses in a *same-different* experiment. End state constraint rankings learned through the GLA yielded only very small differences in the percentage of “different” responses between the monomodally and bimodally-trained grammars (only a 0.1% difference for the simulation reported here). These parameters would need to be further grounded in some phonetic or perceptual basis in order for this model to have predictive power.

### 3. Research Questions and Summary of All “A” Experiments

As stated in the introduction, the main purpose of this study was originally to determine whether Mechanical Turk is a suitable platform for conducting distributional learning experiments. In attempting to



replicate results though, two (arguably) more interesting research questions emerged. Specific research questions in the order they will be presented in are as follows:

- 1) Is there experimental support for the Bias Hypothesis or the Sensitivity Hypothesis? (Sections 4-7)
- 2) Does attention play a role in distributional learning? (Sections 8-9)
- 3) Can Maye and Gerken (2000) be replicated on Mechanical Turk? (Section 10)

The remainder of this section will first present a summary of the designs of the five experiments, referred to in this dissertation as the “A” Experiments, conducted in this chapter. This will be followed by a preview of the conclusions of the research questions stated above.

### 3.1. SUMMARY OF EXPERIMENTAL DESIGNS

The five experiments described in this chapter differ in either critical stimuli used and/or in procedure. Although these differences will be detailed in further sections, this section briefly highlights the main differences in methodology of each experiment. Table 7 provides a summary of these differences.

	<b>Experiment A1</b>	<b>Experiment A2</b>	<b>Experiment A3</b>	<b>Experiment A2-Tone</b>	<b>Experiment A3-Tone</b>
Stimuli	Created by author	Originally used by Maye and Gerken (2001)	Created by author	Originally used by Maye and Gerken (2001)	Created by author
	[gɑ - qɑ] [gæ - qæ] [gɔ̃ - qɔ̃]	[gɑ - qɑ] [gæ - qæ] [gɔ̃ - qɔ̃]	[ɛɑ - ʒɑ]	[gɑ - qɑ] [gæ - qæ] [gɔ̃ - qɔ̃]	[ɛɑ - ʒɑ]
Procedure	---	---	---	Train Check tones included	Train Check tones included
Results: Evidence for distributional learning?	Yes, for Bias Stage	Yes, for Bias Stage	Yes, for Sensitivity Stage	No	No

Table 7. Summary of key differences in all “A” Experiments. Complete summary of is given in Table 16.

Three main experiments were conducted: A1-A3. A pilot experiment related to Experiment A1 seemed to indicate that the inclusion of catch trials during the training phase may play some role in whether we find evidence for distributional learning or not. In order to follow up on these results, Experiments A2 and A3 are each paired with a similar experiment, A2-Tone and A3-Tone respectively, to explore the possible role of attention on distributional learning.

The main difference between Experiments A1-A3 is in their stimuli. Experiment A1 made use of critical stimuli and filler stimuli which were created by the author and which focused on the stop contrast [g-g]. Experiments A2 and A2-Tone made use of critical stimuli and filler stimuli which were originally used by Maye and Gerken (2001) and were subsequently used by Hayes-Harb (2007) which focused on the same [g-g] stop contrast. Experiments A3 and A3-Tone used critical stimuli created by the author, focusing on a fricative contrast ([ε-ʃ]).

### 3.2. SUMMARY OF RESULTS

Two theoretical and two methodological contributions are made in this chapter. The theoretical contributions regard details surrounding distributional learning which have not been previously reported on. They are as follows:

- 1) **Distributional learning does not necessarily affect listeners' sensitivity.** Models suggested by Boersma et al. (2003) and Guenther and Gjaja (1996) directly attribute distributional learning to perceptual warping. These models predict that distributional learning affects listeners' sensitivity. However, this study finds that an increase in sensitivity does not necessarily accompany distributional learning. This chapter proposes an alternate two-stage model of phonetic category learning to explain this result<sup>9</sup> (*See Experiments A1-A3*)

---

<sup>9</sup> Not to be confused with the one- or two-stage models of allophony acquisition discussed in Chapter 4.

- 2) **Attention plays a role in distributional learning.** With the exception of Ong et al. (2015) which will be discussed further in the discussion section, the role of attention has not been noted in previous studies on distributional learning. However, the current study finds evidence to suggest that the level of attention affects whether or not distributional learning occurs (*Compare Experiments A2 and A3 with Experiments A2-Tone and A3-Tone*)

The methodological contributions regard the logistics of carrying out experiments online. These methodological contributions are as follows:

- 3) **Online evidence for distributional learning.** Previous artificial language learning experiments supporting distributional learning have been conducted in lab settings. This chapter describes the first known support for distributional learning in an artificial language learning experiment conducted on an online platform, Mechanical Turk. (*See Experiments A1-A3*)
- 4) **An effect of catch trials during training.** Seemingly minor methodological changes were made to adapt a previous in-lab distributional learning experiment (Maye and Gerken, 2000) to an intended replication study conducted online. Specifically, catch trials were included during training to ensure participants were paying attention. This set of experiments concludes that these catch trials had an effect on participants' responses: experiments which included catch trials failed to replicate Maye and Gerken's results. (*Compare Experiments A2 and A3 with Experiments A2-Tone and A3-Tone*)

The following three sections report the methodology and results of Experiments A1-A3, which differ in the stimuli.

#### **4. Experiment A1**

The original goal of Experiment A1 was to determine whether distributional learning could be replicated through an online platform. Participants were asked to participate only if they (1) had no known history of speech or hearing impairments, (2) were a native speaker of English, (3) had regular access to a computer with an internet connection, and (4) were using a computer able to play audio. Because this experiment

was conducted online rather than face-to-face, only participants using a computer in the United States were allowed to participate to increase the chance that the participant would be a native English speaker. This can be done through MTurk “qualifications,” which are attributes that participants (“Workers,” to use the MTurk terminology) on MTurk can obtain. Qualifications used to screen participants in Experiment A1 were as follows:

- Only Workers who were ages 18-25 were allowed to participate<sup>10</sup>
- Only Workers using a computer in the United States were allowed to participate
- Only Workers who had an approval rating of equal or greater to 90% on all tasks they had completed on MTurk (“HITS”) were allowed to participate
- Only Workers who had at least 50 tasks approved by those putting forth tasks (“Requesters”) were allowed to participate

#### 4.1. STIMULI

Stimuli consisted of critical syllables and filler syllables. Onsets of critical syllables were drawn from three 8-point continua ranging between voiceless unaspirated [g] (*skill*), and prevoiced [g] (*gill*). Continuum points will be referred to as G<sub>1</sub>-G<sub>8</sub>, where G<sub>1</sub> indicates the most [g]-like end of the continuum, and G<sub>8</sub> indicates the most [g̥]-like end. Following Maye and Gerken (2000), each of the three continua differed in following nucleus: [gɑ]-[g̥ɑ], [gæ]-[g̥æ], and [gɪ]-[g̥ɪ]. Stimuli were recorded by the experimenter, a native speaker of English.

Recordings were made in a soundproof booth on an Acer netbook at 44100 Hz using a Logitech H390 USB Headset microphone. Recordings were made in Praat 6.0.29 (Boersma, 2002), software for

---

<sup>10</sup> A previous version of this experiment did not include any restriction on age. This experiment found the unusual result that the Bimodal group responded that pairs were different *less* often than the Monomodal group. This is surprising given the results of Maye and Gerken (2000), Maye and Gerken (2001), and Hayes-Harb (2007), who find that the group trained on a bimodal distribution of critical phones is significantly *more* likely to respond that critical pairs are different compared to the group trained on a monomodal distribution of critical phones. Although differing in a few other respects to the design of the experiments discussed here, it was decided that only 18-25 year olds should be tested in future experiments, as it was assumed that most participants in previous distributional learning experiments were undergraduate students. Results of this experiment will not be discussed here, but can be found in Moeng (2017).

speech analysis, synthesis, and manipulation. Before any manipulations were performed, all stimuli (both critical and filler) were scaled to a peak intensity of 72 dB in Praat. The experimenter recorded tokens of [sk-] and [g-] followed by each of the three context vowels [a æ ə] and removed the [s] portion from the [sk-] initial syllables. These formed the end points of each of the [g]-[g] continua. Prevoicing was then removed from the [g-] syllables. All cuts were made where the waveform crossed 0 Hz to avoid clicks and other unnatural non-speech sounds when splicing sounds together. All splicing was done in Praat. Each of the three pairs of endpoints ([gɑ gæ gə] from [sk-] syllables with the [s] portion removed, and [gɑ gæ gə] with the pre-voicing removed) were then input into TANDEM-STRAIGHT (details of the process TANDEM-STRAIGHT uses to create continua can be found in Kawahara et al. (2008)), which is a piece of software which creates natural-sounding continua between two sounds. TANDEM-STRAIGHT allows the user to mark any number of landmarks on one spectrogram (for example, the beginning of the steady state of the vowel, the onset of voicing, etc.) that corresponds with a similar landmark on another spectrogram, so that durations between landmarks can be stretched or compressed in the intervening continuum points. It has also been used in other linguistic studies to create continua (for example, Noguchi, 2016; McAuliffe and Babel, 2016). TANDEM-STRAIGHT returned 6 intermediate stimuli, for a total of 8 continuum points including the endpoints. Following this, the prevoicing which had been removed from the [g-] portion of the [gæ] token (which was 140 ms in length) was shortened into 8 equal prevoicing portions ranging from 0-140 ms in length (0, 20, 40... 140). These prevoicing portions were then spliced back onto each of the continuum points (for all three continua), with the 140 ms prevoicing portion being spliced onto the [g]-most end (G<sub>1</sub>), the 20 ms prevoicing portion being spliced onto the penultimate of the [g]-most end (G<sub>7</sub>), and the [g]-most end (G<sub>8</sub>) having no prevoicing spliced on. (Therefore, all three completed continua began with varying lengths of prevoicing, all manipulated from the [gæ] recording.)

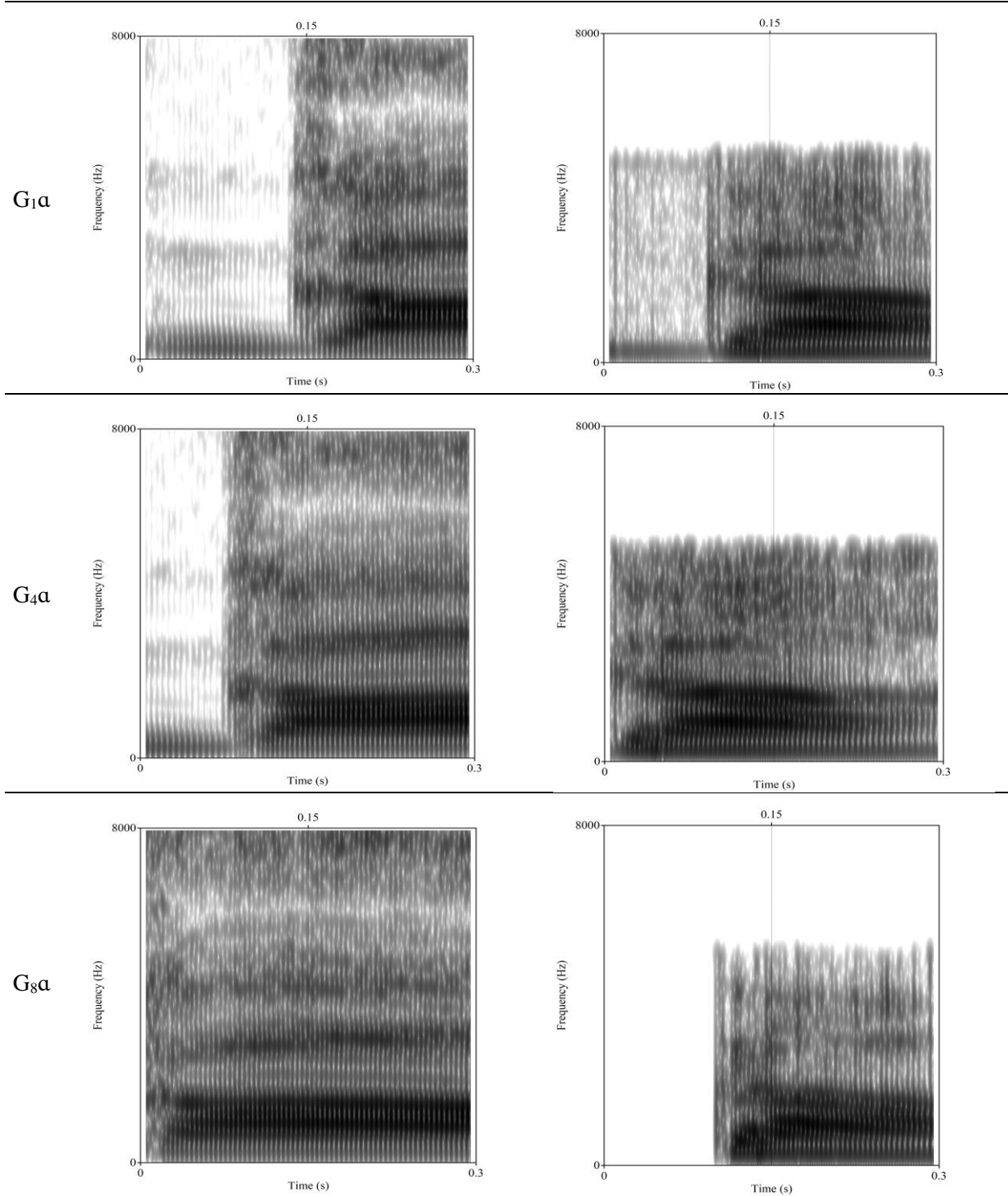


Figure 23. First 300 ms of spectrograms of G<sub>1a</sub>, G<sub>4a</sub>, and G<sub>8a</sub> for stimuli created by the author (left), and for stimuli created by Jessica Maye and LouAnn Gerken (right).

All stimuli were judged by the author, a native speaker of English, to sound natural. For visual reference, Figure 23 (left-most column) shows the first 300 ms of the spectrograms of the continuum points  $G_{1a}$ ,  $G_{4a}$ , and  $G_{8a}$ , as created by the author. For comparison, the first 300 ms of the spectrograms of  $G_{1a}$ ,  $G_{4a}$ , and  $G_{8a}$  stimuli originally used by Maye and Gerken (2001) and also used by Hayes-Harb (2007) are shown on the right. Some notable differences between the stimuli created by the author and the stimuli created by Maye and Gerken (2001) are as follows: 1) the stimuli created by the author (left) were about twice as long in duration compared to the stimuli created by Maye and Gerken (for example, 768 ms vs. 468 ms for  $G_{1a}$ ), and 2) the stimuli created by Maye and Gerken have a cutoff frequency of 5000 Hz, whereas the stimuli created by the author have a cutoff frequency of 22050 Hz. It was not believed that these differences would affect performance, and so were left in.

#### 4.2. PROCEDURE

Experiment A1 consisted of 5 parts, summarized in Table 8 and described in this section.

<b>Procedure</b>	<b>Abbreviated directions</b>	<b>Sample trial stimuli</b>
1. Sound check	Press “1” or “2” to indicate beeps heard	<i>::one tone::</i> <i>::two tones::</i>
2. Practice test	Are these two words the same or different?	<i>sheep vs. ship</i> <i>sheep<sub>1</sub> vs. sheep<sub>2</sub></i>
3. Training	Listen carefully	<i>G<sub>1a</sub></i> <i>ma</i>
4. Test	Are these two words the same or different?	<i>G<sub>1a</sub> vs. G<sub>8a</sub></i> <i>G<sub>1a</sub> vs. G<sub>1a</sub></i>
5. Questionnaire	Please provide background information (responses will not affect payment)	

Table 8. Summary of procedure in Experiment A1.

At the beginning of the experiment, participants were given the following instructions:

*[Page 1]*

*Just a few things to keep in mind before you begin!*  
*Please wear headphones for the duration of this experiment...*  
*Please do not write any words down while taking this experiment...*  
*...And please do not click on your browser's back or refresh button.*

[Page 2]

*A few more things to keep in mind before you begin...*

*As this is a scientific experiment, it is important that you devote your **full attention** to this HIT<sup>11</sup>! There will be a number of checks in place to ensure that you are paying attention to the task at hand, and you may not be paid if you do not pass these checks. Please do not do other tasks or shrink the browser window, and please do not remove your headphones.*

*If you are unable to devote your full attention to this HIT, which is expected to take about 30 minutes, **please do not do this HIT.***

Participants were then directed to a Sound Check, the purpose of which was to 1) ensure participants were wearing headphones, and 2) encourage participants to pay attention and not give random responses. The Sound Check consisted of 3 one-tone tokens and 3 two-tone tokens, presented in random order. Participants were instructed to press the “1” or “2” keys if they heard one of these “beep” tokens, to indicate how many “beeps” they had heard. Tones were chosen to be at a low enough frequency that most computer speakers would not pick up on the sound (50 Hz), thereby testing whether participants were wearing headphones or not. Tones were 340 ms long, and were spaced by 140 ms of silence for the two-tone tokens. Participants were excluded from analysis if they failed to answer 5 out of 6 of these trials correctly. 12 participants in Experiment A1 failed to meet this criterion.

Following the Sound Check, participants were directed to a Practice Test phase. During the Practice Test phase, participants were given the following instructions:

[Page 1]

*In this English practice test, you will hear two words in English, and will be asked if they are repetitions of the same word, or if they are two different words*

*If you think they are repetitions of the same word, press the “S” key for “Same”.*

*If you think they are different words, press the “D” key for “Different”.*

---

<sup>11</sup> A “HIT” (Human Intelligence Task) is a term used in Mechanical Turk, referring to a task that a participant (“Worker”) can complete.



Participants were then presented with pairs of English words produced by the same speaker that were either Same Pairs, or Different Pairs. Same Pairs consisted of repetitions of the same word that were different enough to be distinguished as different tokens (e.g. *lock*<sub>1</sub> vs. *lock*<sub>2</sub>). Different Pairs consisted of English minimal pairs (e.g. *lock* vs. *rock*, *desk* vs. *disk*). Participants were asked to press the “S” key if the pairs of words that they heard were the “same” word, or the “D” key if they were “different” words. Pairs were separated by 1 second, and participants were given 10 seconds to respond before the next pair was played. Responses were scored as “correct” if participants responded “different” on Different Pairs and “same” on Same Pairs. Participants who answered fewer than 5/8 correct on the Practice Test were excluded. No participants in Experiment A1 failed to meet this criterion.

During the Train phase, participants heard a monomodal or bimodal distribution of phones, depending on which condition they were in. The Bimodal group heard a bimodal frequency of phones of the frequencies shown in the dotted line in Figure 2, and the Monomodal group heard a monomodal frequency distribution of phones of the frequencies shown in the solid line in Figure 2. These frequency values follow those used by Maye and Gerken (2000). This resulted in 16 critical tokens from each of the three continua (1+1+2+4+4+2+1+1, or 1+4+2+1+1+2+4+1). In addition, three recordings of 8 filler syllables ([fa], [fæ], [tɛ], [tej], [mæ], [næ], [sɛ], and [zɛ]) were made. Each of these 24 filler tokens were repeated twice during each Train repetition. Tokens within a Train repetition were presented in random order.

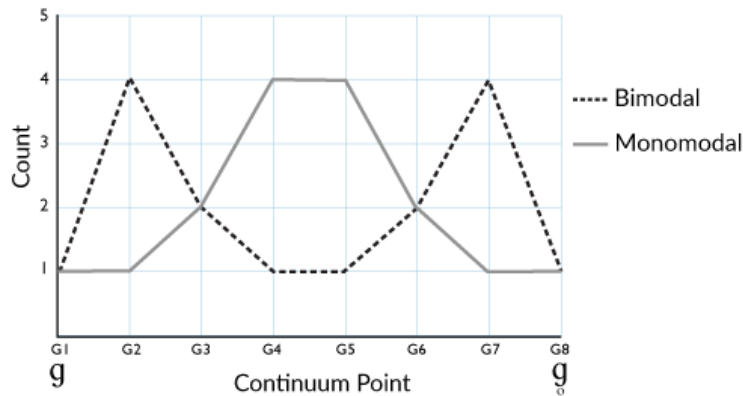


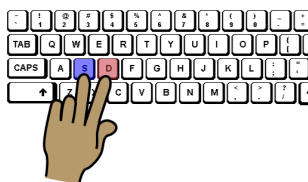
Figure 2. Familiarization frequency of critical stimuli for the Bimodal (dashed line) and Monomodal (solid line) groups during the Train phase.

Each Train repetition was repeated 4 times, resulting in a total of 192 fillers, and 192 critical tokens.

Following the Train phase, participants were directed to a Test phase. The Test phase was similar to the Practice Test phase, except participants were given pairs of words they had heard in the artificial language they had heard during the Train phase. Again, participants were given pairs of syllables that were either Same Pairs, or Different Pairs. Critical Same Pairs consisted of repetitions of the same exact token (i.e.  $G_{1a}$  vs.  $G_{1a}$ ; or  $G_{8ae}$  vs.  $G_{8ae}$ ), while filler Same Pairs consisted of different tokens of the same syllable ( $[s\epsilon]_1$  vs.  $[s\epsilon]_2$ ). Filler Same Pairs were judged by the experimenter to sound different enough to be distinguished as separate tokens. Critical Different Pairs consisted of pairs that occurred on opposite ends of the 8-point continuum (i.e.  $G_{1a}$  vs.  $G_{8a}$ ), whereas filler Different Pairs consisted of different pairs of syllables ( $[s\epsilon]$  vs.  $[z\epsilon]$ ). During the Test phase, participants were shown the following instructions:

*This next part will be similar to the practice testing you did earlier in English, **but this time it will ask you about the made-up language that you just heard.***

*Like before, please place one finger over the "S" key and another key over the "D" keys on your keyboard, as shown below.*



*Like before, if you think they are repetitions of the same word, press the "S" key for "Same".*

*If you think they are different words, press the "D" key for "Different".*

Words in a pair were separated by 1 second, and participants were given 10 seconds to respond before the next pair was played.

Each Test repetition consisted of 4 critical pairs (2 Same Pairs and 2 Different Pairs), for each of the three vowel contexts, resulting in 12 critical pairs. Each Test repetition also contained 4 filler Same Pairs and 4 filler Different Pairs. There were two repetitions of each Test phase resulting in a total of 24 critical pairs and 16 filler pairs. Participants were instructed to press the "S" key if the pairs of words that they heard were the "same" word, or the "D" key if they were "different" words. Pairs were separated by 1 second, and participants were given 10 seconds to respond before the next pair was played.

Following the Test phase, participants were directed to a questionnaire. The questionnaire included questions regarding language background, a question regarding English fluency, and a question regarding having a history of a speech or hearing disorder. Participants were told that their responses would not negatively affect them and to please answer truthfully. Participants were excluded from analysis if they reported not being a native speaker of English (no participants in Experiment A1 were excluded from analysis for this reason), or if they reported having a history of a speech or hearing disorder (1 participant in Experiment A1 was excluded from analysis for this reason). A copy of the questionnaire is included in the Appendix.

Participants were placed randomly into one of two conditions: a Monomodal group or a Bimodal group. In total, 13 participants were rejected from analysis (some for multiple reasons), leaving 27 in the Bimodal group and 34 in the Monomodal group.

### 4.3. ANALYSIS

This section will first describe the model used to analyze results for all “A” Experiments, and then explain how the results of the fitted model will be interpreted. Section 4.4 will present the results of the fitted model for Experiment A1.

#### 4.3.1 Model used in analysis

The regression formula described below will model one dependent variable, **Response**, with two fixed effects: (1) **Distribution**, a between-subject, within-item factor consisting of two levels {*bimodal*, *monomodal*} and (2) **PairType**, a within-subject, between-item factor consisting of two levels {*same*, *diff*}.

This will be done separately for *critical* and *filler* trials. The dependent variable **Response** consists of two levels, *s* and *d*, where *s* corresponds to a participant response of “same” during the Test phase, and where *d* corresponds to a participant response of “different” during the Test phase. Random effects for **Subject** and **Item** will also be included in the model described below. All variables are summarized in Table 9.

Variable type	Effect type	Factor name	Factor type	Level names	Description
Independent variables	Fixed effects	Distribution	Between-subject, within-item	<i>bimodal</i> <i>monomodal</i>	Distribution type received by the participant during Train phase
		PairType	Within-subject, between-item	<i>same</i> <i>diff</i>	Type of pair presented during Test phase (Same Pair or Different Pair)
	Random effects	Subject			Each individual participant (coded by ID)
		Item			Each individual item presented during the Test phase
Dependent variables		Response		<i>s</i> <i>d</i>	Response given by participant

Table 9. Variables used in regression analysis for “A” Experiments.

All statistical tests were completed in R (R Core Team, 2014), using the `glmer` function from the `lme4` package (Bates et al. 2015) to fit a generalized linear mixed-effects model (GLMM) with a logistic link function (“mixed logit model”). Significance was set at a level of  $p < 0.05$ . Two separate regressions were conducted to compare the effect of Distribution and PairType on Response: one for *critical* items, and one for *filler* items. This dissertation follows suggestions made by Clark (1973) for variables to include as

random effects, and suggestions made by Barr et al. (2013) for how to translate random effects used in an ANOVA into the random effects structure of the formula used to fit a regression. Clark (1973) argues that items, and not just subjects, must be accounted for as random effects in an ANOVA. That is, not only are individual subjects more or less inclined to respond one way or another, but individual test items also bring their own idiosyncratic individual behavior, which may not be entirely representative of the entire lexicon, to the experiment. The inclusion of subject and item random effects increases the generalizability of the results of this particular experiment to other subjects, as well as to other test items.

This dissertation also follows suggestions made by Barr et al. (2013) for the random effects structure of the formula fitted in the regression. Barr et al. argue that if a factor *would* be a between-subject factor in an ANOVA, it is sufficient to include only a random intercept by subject into the random effects structure in a regression. Likewise, if a factor would be a between-item factor in an ANOVA, it is sufficient to include only a random intercept by item into the random effects structure. However, if a factor would be a within-subject factor in an ANOVA, both a random intercept by subject as well as a random slope by subject are necessary in a regression model. Likewise, if a factor would be a within-item factor in an ANOVA, both a random intercept by item as well as a random slope by item are necessary to include in the random effects structure in a regression model. In a follow-up paper, Barr (2013) claims that a random effects structure is unnecessary to account for interactions between a between-subject factor and a within-subject factor.

Since the current design consists of one between-subject within-item factor (Distribution) as well as one within-subject between-item factor (PairType), the formula used in the regression is as follows:

$$(5) \quad \text{Response} \sim \text{Distribution*PairType} + (1+\text{PairType}|\text{Subject}) + (1+\text{Distribution}|\text{Item})$$

The formula in (5) tests for the effect on participant Response of the interaction between Distribution and PairType, for the simple effect of Distribution within the reference of PairType (*diff*), and for the simple

effect of PairType within the reference level of Distribution (*bimodal*). A random slope by subject is included for PairType and a random slope by item is included for Distribution.

#### 4.3.2 Model interpretation

The results of the mixed logit model are interpreted in the following way, as justified in Chapter 2: a significant main effect of Distribution *without* a significant interaction between Distribution and PairType is interpreted as evidence for a difference in **response bias** (see Section 2.3) between the two groups of participants. That is, if participants trained on Distribution X have a *greater* bias towards a “different” response compared to participants trained on Distribution Y, we would expect to see participants trained on Distribution X to respond “different” more often than participants trained on Distribution Y for all trials, regardless of whether or not the trial consisted of a Same Pair (e.g. G<sub>1</sub>α vs. G<sub>1</sub>α) or a Different Pair (e.g. G<sub>1</sub>α vs. G<sub>8</sub>α). We expect to see this in the results of a fitted model as a significant main effect of Distribution, but *not* as a significant interaction between Distribution and PairType since, in this scenario, the effect of PairType would not differ depending on which Distribution participants were exposed to.

An interaction between Distribution and PairType will be interpreted as a difference between groups of participants in **sensitivity** to the slight distinction between G<sub>1</sub> and G<sub>8</sub>. That is, if participants trained on Distribution X are more sensitive to the acoustic differences between G<sub>1</sub> and G<sub>8</sub> compared to participants trained on Distribution Y, we would expect participants trained on Distribution X to respond “different” more often than participants trained on Distribution Y for critical Different Pairs, but to also respond “different” *less* often for critical Same Pairs, resulting in a significant interaction between Distribution and PairType.

#### 4.4. RESULTS

A generalized linear mixed model with a logit link function (GLMM) was fitted to the formula in (5), where the reference cell was *bimodal diff*. Summaries of the fixed effects in the mixed logit model with treatment coding are shown in Table 10. No interaction between Distribution and PairType was found for

critical trials ( $p = 0.650$ ), but an interaction was found for filler trials ( $p = 0.047$ ). To test for a main effect of Distribution on Response, a planned contrast analysis was performed in the context of the overall model using the `glht` function from the `multcomp` package (Hothorn et al., 2017) in R. This planned contrast revealed a main effect of Distribution for critical trials ( $p = 0.049$ ), with the Bimodal group having greater odds of responding  $d$  than the Monomodal group. No main effect was found for filler trials ( $p = 0.817$ ).

Predictor	Coefficient	SE	Wald Z	$p$
<b>CRITICAL</b>				
(Intercept)	-1.502	0.405	-3.707	<0.001 ***
Distribution= <i>monomodal</i>	-0.990	0.434	-2.280	0.023 *
PairType= <i>same</i>	-4.174	1.652	-2.526	0.012 *
Interaction= <i>monomodal &amp; same</i>	-0.513	1.130	-0.454	0.650
<b>FILLER</b>				
(Intercept)	2.834	0.517	5.486	<0.001 ***
Distribution= <i>monomodal</i>	0.769	0.605	1.271	0.204
PairType= <i>same</i>	-4.169	0.709	-5.881	<0.001 ***
Interaction= <i>monomodal &amp; same</i>	-1.743	0.877	-1.988	0.047 *

Table 10. Summary of fixed effects in the mixed logit model in Experiment A1.



Figure 24. Log odds of participants responding  $d$  for critical trials (left) and filler trials (right) in Experiment A1. Error bars indicate standard error.

#### 4.5. DISCUSSION

If “distributional learning” is taken to only refer to greater sensitivity for bimodally-trained learners compared to monomodally-trained learners, then Experiment A1 fails to find evidence for distributional learning. If, however, the definition of “distributional learning” is more widely-defined to include a greater bias towards a “different” response for bimodally-trained learners compared to monomodally-trained learners, then Experiment A1 is the first indication known to the author that distributional learning can be replicated through an online platform such as MTurk. Further, Experiment A1 appears to support the Bias Hypothesis of distributional learning.

Unexpectedly, a significant interaction between Distribution and PairType was found for the filler trials, with the bimodal group having lesser sensitivity to filler trials compared to the monomodal group. This and other unexpected results will be discussed in Section 11.

In order to determine if these results were particular to the stimuli of this experiment, Experiment A1 was followed by Experiment A2. The goal of Experiment A2 is to replicate the results of Experiment A1 with stimuli from Maye and Gerken (2001).

### **5. Experiment A2**

The goal of Experiment A2 was to replicate the results of Experiment A1 using stimuli which have been used in past studies. Experiment A2 followed the methodology of Maye and Gerken (2001) and Hayes-Harb (2007) as closely as possible, using stimuli obtained from Maye and Gerken. The procedure also closely followed that used in Maye and Gerken (2001), with the small addition of the Sound Check task (identical to that used in Experiment A1) preceding the experiment.<sup>12</sup> The following exclusion criteria were used:

---

<sup>12</sup> Many thanks to Rachel Hayes-Harb, LouAnn Gerken, and Jessica Maye for sending me and allowing me to use their stimuli.



- Fewer than 5/6 on the pre-experiment Sound Check (7 excluded)
- Fewer than 5/8 correct on the Practice Test (1 excluded)
- Reported not being a native speaker of English (0 excluded)
- Reported a history of a speech or hearing disorder (1 excluded)

In total, 7 participants were rejected from analysis from Experiment A2, leaving 21 in the Bimodal group and 27 in the Monomodal group.

### 5.1. PROCEDURE

The procedure of Experiment A2 was identical to that followed by both Maye and Gerken (2000) and the phonetic learning part of the experiment conducted by Hayes-Harb (2007), but was preceded by the Sound Check task described in Experiment A1. As was the case for Experiment A1, this experiment consisted of a Sound Check, Practice Test phase, a Train phase, and a Test phase, followed by a Questionnaire. Each Train repetition consisted of 16 critical tokens for each of the three vowel contexts, two repetitions of four separate tokens of 6 fillers [mə mə mə la læ lə]. Each Train repetition was repeated 4 times, resulting in a total of 192 fillers, 192 critical tokens. As noted earlier, stimuli are those used by Maye and Gerken (2001) as well as Hayes-Harb (2007). Examples of select continuum points can be found in Figure 23 on page 67. The Test phase consisted of 2 critical Same Pairs and 2 critical Different Pairs for each of the 3 vowel contexts resulting in 12 critical pairs, as well as 2 filler Same Pairs and 2 filler Different Pairs for each of the 3 vowel contexts, resulting in 12 filler pairs.

### 5.2. RESULTS

Again, a generalized linear mixed model with a logit link function (GLMM) was fitted to the formula in (5), with the reference cell being *bimodal diff*. Summaries of the fixed effects in the mixed logit model with treatment coding for Experiment A2 are shown in Table 11. For Experiment A2, no interaction between Distribution and PairType was found for critical trials ( $p = 0.114$ ), or for filler trials ( $p = 0.228$ ).

Predictor	Coefficient	SE	Wald Z	p
<b>CRITICAL</b>				
(Intercept)	-2.152	0.318	-6.773	<0.001 ***
Distribution= <i>monomodal</i>	-1.420	0.485	-2.927	0.003 **
PairType= <i>same</i>	-1.255	0.471	-2.667	0.008 **
Interaction= <i>monomodal &amp; same</i>	0.982	0.622	1.579	0.114
<b>FILLER</b>				
(Intercept)	2.977	0.471	6.316	<0.001 ***
Distribution= <i>monomodal</i>	0.529	0.568	0.932	0.351
PairType= <i>same</i>	-5.720	0.673	-8.503	<0.001 ***
Interaction= <i>monomodal &amp; same</i>	-1.001	0.830	-1.206	0.228

Table 11. Summary of fixed effects in the mixed logit model in Experiment A2.

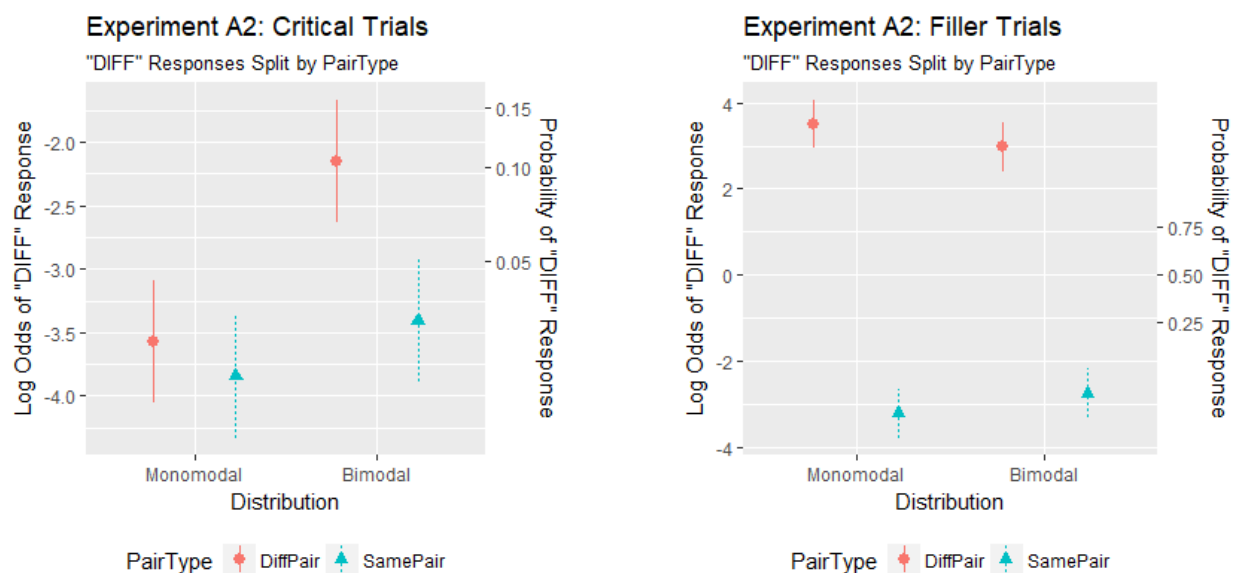


Figure 25. Log odds of participants responding *d* for critical trials (left) and filler trials (right) in Experiment A2. Error bars indicate standard error.

To test for a main effect of Distribution on Response, a planned contrast analysis was performed on the fitted GLMM using the `glht` function from the `multcomp` package (Hothorn et al., 2017) in R. The planned contrast revealed a main effect of Distribution was found for critical trials ( $p = 0.033$ ), with the Bimodal group having greater odds of responding *d* than the Monomodal group. No main effect was found for filler trials ( $p = 0.934$ ).

### 5.3. DISCUSSION

As with Experiment A1, no significant interaction between Distribution and PairType was found. However, there was a significant main effect of Distribution on participant response, with participants in the Bimodal group having greater (log-)odds of responding “different” than participants in the Monomodal group. Therefore bimodally-trained participants exhibit a greater bias towards a “different” response compared to monomodally-trained ones in both Experiment A1 and Experiment A2. Again, Experiment A2 appears to find support for the Bias Hypothesis of distributional learning.

### 6. Experiment A3

The two previous experiments found evidence that bimodally-trained participants have a greater **bias** towards a “different” response compared to monomodally-trained participants. However, previous studies claim to find greater **sensitivity** in bimodally-trained participants compared to monomodally-trained ones. The goal of Experiment A3 was to determine if using different critical stimuli resulted in greater sensitivity for bimodally-trained participants. It was believed that the stimuli used in Experiments A1 and A2 did not lead to a difference in sensitivity among conditions because they were perceptually very similar. It was thought that using critical stimuli which ranged between more perceptually distinct endpoints might lead to evidence that bimodal training causes an increase in sensitivity compared to monomodal training. Therefore this experiment will use a contrast used by Noguchi (2016), which tests a fricative distinction [ɛɑ-ʂɑ].

Although the original intent was to choose critical endpoint stimuli which are more perceptually distinct from one another, it should be noted that [g-g] and [ɛ-ʂ] differ in more than just their perceptual distinctness. This issue is discussed further in the discussion section.

Following Experiments A1 and A2, Experiment A3 used the following criteria to exclude participants from analysis:

- Fewer than 5/6 on the pre-experiment Sound Check (28 excluded<sup>13</sup>)
- Fewer than 5/8 correct on the Practice Test (4 excluded)
- Reported not being a native speaker of English (1 excluded)
- Reported a history of a speech or hearing disorder (0 excluded)

As detailed later, critical stimuli fall along a continuum between [ɛɑ] – [ʂɑ], so unlike the previous experiments, this pair of experiments also excluded participants who had experience with more than one voiceless post-alveolar fricative as phonemes, following Noguchi (2016) who also used this contrast as critical stimuli. Before the experiment, participants were asked to not participate if they had studied or had experience with German, Mandarin, Japanese, or Russian. In a following questionnaire, they were asked what experience they had with other languages and were told that their responses would not affect payment. 3 participants in Experiment A3-2 were excluded for reporting experience with a language with more than one voiceless post-alveolar fricative. In total, 32 participants were rejected from analysis from Experiment A3 (some for multiple reasons), leaving 22 in the Bimodal group and 27 in the Monomodal group.

## 6.1. STIMULI

Stimuli consisted of critical syllables and filler syllables. Following Noguchi (2016), onsets of critical syllables were drawn from an 8-point continuum ranging between an alveopalatal fricative [ɛ] to a retroflex fricative [ʂ]. Continuum points will be referred to as S<sub>1</sub>-S<sub>8</sub>, where S<sub>1</sub> indicates the most [ɛ]-like end of the continuum, and S<sub>8</sub> indicates the most [ʂ]-like end. Although previous experiments followed critical onsets with three different rimes ([ɑ æ ə]), the current experiment follows Noguchi (2016) and all onsets are only followed by [ɑ]. Filler syllables consisted of the syllables [ta t<sup>h</sup>ɑ fa ha].

---

<sup>13</sup> A large number of participants were excluded for this reason. I believe this is because the initial instructions that MTurkers received when determining whether they want to participate in this experiment differed from the initial instructions in the A1 and A2 experiments. In the initial instructions of the A1 and A2 experiments, prospective participants were directed to a page where they could play the one or two tones. These prospective participants were told that if they could not hear the tones played on this page, they should not take the experiment. The A3 experiments failed to include this.

All recordings, filtering, and splicing were done in Praat (version 6.0.29, Boersma, 2002), software for speech analysis, synthesis, and manipulation. Stimuli were recorded by the experimenter, a native speaker of English and heritage speaker of Mandarin. Recordings were made in a soundproof booth on an HP Spectre laptop at 44100 Hz using an ATR2500-USB Audio Technica microphone. Before manipulations were made, all recordings were high-pass filtered for frequencies equal to or below 200 Hz. All cuts were made where the waveform crossed 0 Hz to avoid clicks and other unnatural non-speech sounds when splicing sounds together.

For critical syllables, tokens of the two endpoints of the target continuum [ɛɑ] and [ʂɑ] were recorded. Recordings were made such that test syllables were preceded by a dummy syllable [ɑ] (e.g. [ɑ ɛɑ]), as these same test syllables would be used in Experiment B, reported on in Chapter 4. However, test syllables [ɛɑ] and [ʂɑ] sounded like they had been produced in isolation. The fricative portions of the test syllables [ɛɑ] and [ʂɑ] ([ɛ] and [ʂ]) and vowel portions ([ɑ]) were isolated. The middle 160 ms of each fricative was extracted using a parabolic windowing function. The mean intensity of each fricative was adjusted to 60 dB. To create the fricative portion of the 8-point continuum, the endpoint fricatives were overlapped in varying amounts, with the second point of the continuum consisting of 6/7ths of the [ɛ] token and 1/7th of the [ʂ] token, the third point of the continuum consisting of 5/7ths of the [ɛ] token and 2/7th of the [ʂ] token, etc. The continuum between vowels was created by using TANDEM-STRAIGHT, software which creates natural-sounding continua between two sounds. The vowel spliced from [ɛɑ] and the vowel spliced from [ʂɑ] were used as input into TANDEM-STRAIGHT (Kawahara et al., 2008). TANDEM-STRAIGHT allows the user to mark any number of landmarks on one spectrogram (for example, the beginning of the steady state of the vowel, the onset of voicing, etc.) that corresponds to a similar landmark on another spectrogram, so that durations between landmarks can be stretched or compressed in the generated continuum points. TANDEM-STRAIGHT returned 6 intermediate stimuli, for a total of 8 continuum points including the endpoints. All vowel continuum points were then scaled to have a mean intensity of 72 dB. Following this, each of the 8 fricative sounds were spliced onto their corresponding 8

vowel sounds, creating an 8-point continuum between [eɑ] and [ɤɑ]. Each of these syllables was scaled to have an average intensity of 74 dB. Critical syllables will be referred to as S<sub>1a</sub>-S<sub>8a</sub>, where S<sub>1a</sub> refers to the most [eɑ]-like end, and S<sub>8a</sub> refers to the most [ɤɑ]-like end. Examples of critical syllables S<sub>1a</sub> and S<sub>8a</sub> are shown in Figure 26.

Experiment A3 Stimuli

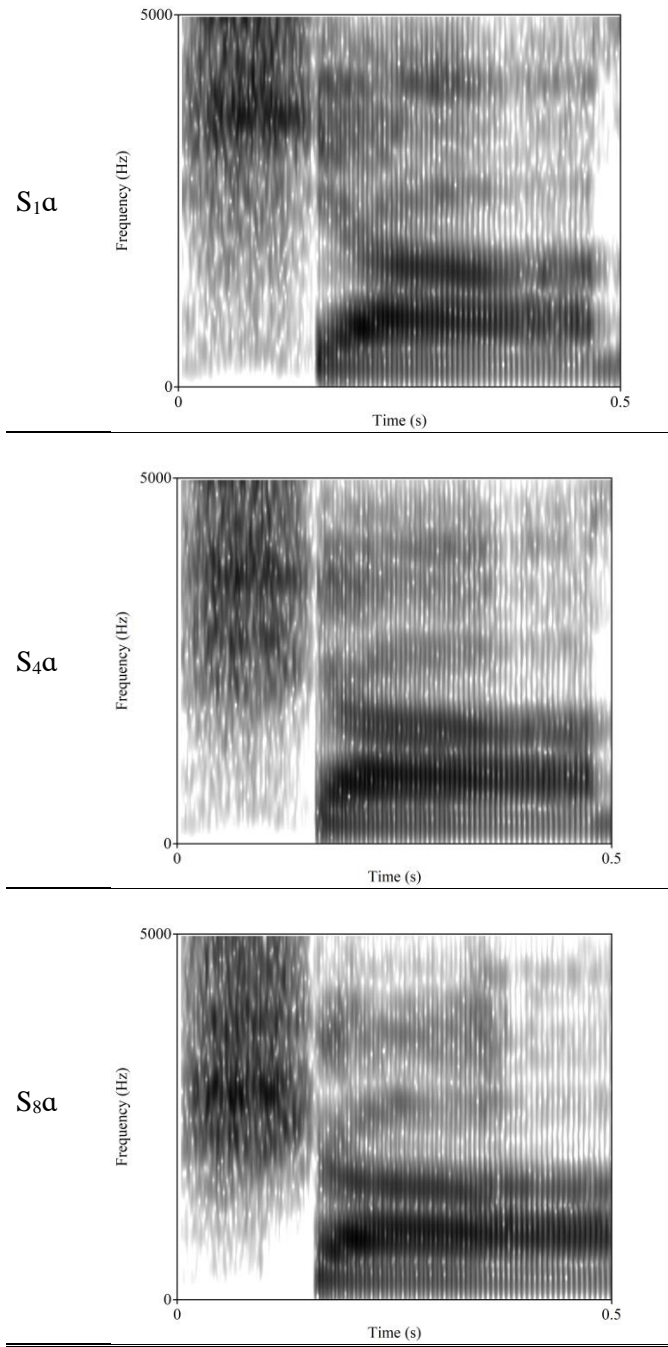


Figure 26. First 500 ms of critical syllables S<sub>1a</sub> (top), S<sub>4a</sub> (middle), and S<sub>8a</sub> (bottom).

## 6.2. PROCEDURE

Again, the procedure consisted of a Sound Check, Practice phase, a Train phase, and a Test phase, followed by a Questionnaire. The Sound Check, Practice phase, and all instructions were identical to those used in previous “A” Experiments. As with previous “A” Experiments, participants heard a monomodal or bimodal distribution of the 8-point continuum of critical phones, depending on which condition they were in. Rather than hearing three different rime contexts though, participants heard three repetitions of each distribution of critical  $S_{1-8a}$  tokens per block. In addition, 4 recordings of 4 filler syllables ([t<sup>h</sup>a], [ta], [fa], [ha]) were made. Each of these 16 filler tokens were repeated 3 times during each train block. Each block was repeated 4 times, resulting in a total of 192 critical tokens, and 192 filler tokens. The Train phase lasted for about 10 minutes.

As with the previous “A” Experiments, the Test phase presented participants with pairs of syllables that were either Same Pairs, or Different Pairs. Same Pairs consisted of repetitions of the same exact token for critical tokens (e.g.  $S_{7a}$  vs.  $S_{7a}$ ), or different tokens for filler tokens (e.g. [fa]<sub>1</sub> vs. [fa]<sub>2</sub>). Filler Same Pairs were judged by the experimenter to sound different enough to be distinguished as separate tokens. Critical Different Pairs consisted of pairs that occurred on opposite ends of the 8-point continuum, for critical tokens (i.e.  $S_{1a}$  vs.  $S_{8a}$ ). The Test phase consisted of 12 critical Same Pairs, 12 critical Different Pairs, 12 filler Same Pairs, and 12 filler Different Pairs.

## 6.3. RESULTS

A generalized linear mixed model with a logit link function (GLMM) was fitted to the formula in (5), with the reference cell being *bimodal diff*. Summaries of the fixed effects in the mixed logit model with treatment coding for Experiment A3 are shown in Table 12. An interaction between Distribution and PairType was found for critical trials ( $p = 0.037$ ). Surprisingly, an interaction between Distribution and PairType was found for filler trials as well ( $p = 0.009$ ). Results are shown in Figure 27.



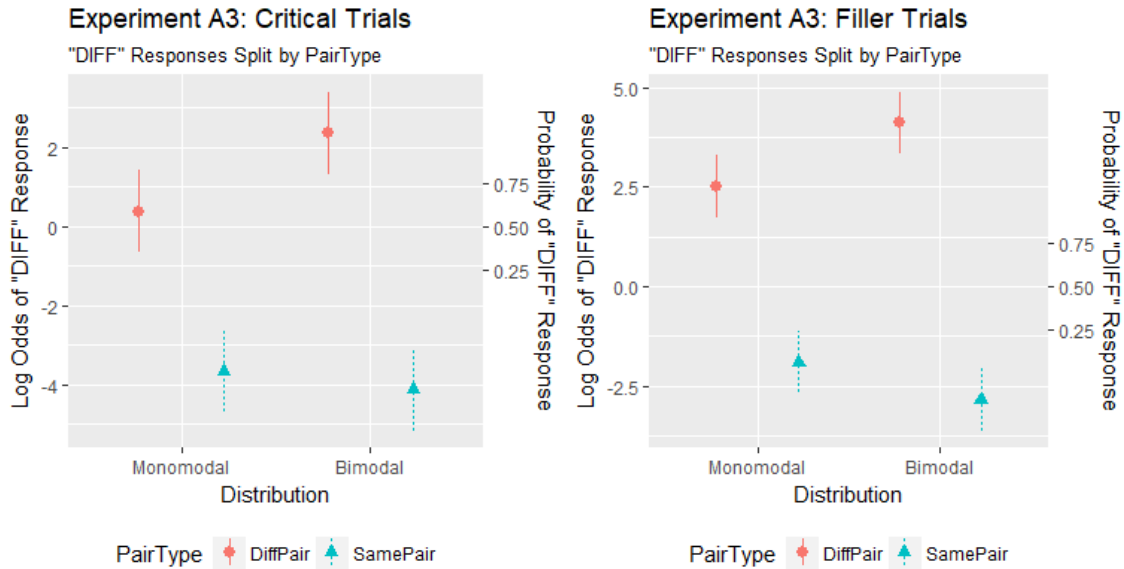


Figure 27. Log odds of participants responding *d* for critical trials (left) and filler trials (right) in Experiment A2. Error bars indicate standard error.

Predictor	Coefficient	SE	Wald Z	<i>p</i>
<b>CRITICAL TRIALS</b>				
(Intercept)	2.365	0.809	2.922	0.003 **
Distribution= <i>monomodal</i>	-1.978	1.040	-1.903	0.057
PairType= <i>same</i>	-6.489	0.963	-6.738	<0.001 ***
Interaction= <i>monomodal &amp; same</i>	2.439	1.166	2.091	0.037 *
<b>FILLER TRIALS</b>				
(Intercept)	4.112	0.759	5.415	<0.001 ***
Distribution= <i>monomodal</i>	-1.599	0.780	-2.052	0.040 *
PairType= <i>same</i>	-6.957	0.944	-7.371	<0.001 ***
Interaction= <i>monomodal &amp; same</i>	2.549	0.978	2.606	0.009 **

Table 12. Summary of fixed effects in the mixed logit model in Experiment A3.

To test for a main effect of Distribution on Response, a planned contrast analysis was performed on the fitted GLMM using the `glht` function from the `multcomp` package (Hothorn et al., 2017) in R. The planned contrast revealed no main effect of Distribution for critical trials ( $p = 0.229$ ) or filler trials ( $p = 0.474$ ).<sup>14</sup>

<sup>14</sup> This test is included for symmetry with Experiments A1 and A2, but note that this dissertation would not have interpreted a significant main effect of Distribution, since a significant interaction between Distribution and PairType had already been found (see Chapter 2 for reasoning).

#### 6.4. DISCUSSION

Although an unexpected significant interaction between Distribution and PairType was found for filler trials (discussed further in Section 11), this section concludes that the exposure length (192 critical trials) and stimuli used in Experiment A3 was able to induce greater sensitivity in a bimodally-trained group of participants compared to a monomodally-trained group of participants, as a significant interaction between Distribution and PairType was found for critical trials. Experiment A3 appears to be the first evidence that **sensitivity** changes through distributional learning, and not only bias changes, can be found through an online platform. While Experiments A1 and A2 found support for the Bias Hypothesis, the results of Experiment A3 appear to support the Sensitivity Hypothesis.

As noted earlier, the original intent in choosing the [ɛ-ʒ] contrast was to choose critical endpoint stimuli which are more perceptually distinct from one another, in the hopes that this would result in a difference in sensitivity between conditions. However, [g-g] and [ɛ-ʒ] differ in more than just their perceptual distinctness. Specifically, it could be argued that [g] and [g] belong to separate phonemes for English speakers ([g] being an allophone of /g/ and [g] being an allophone of /k/), or it could be argued that English speakers categorize syllable-initial [g] and [g] as free variants of a single phoneme /g/. On the other hand, [ɛ] and [ʒ] could be argued to be both categorized by an English listener as a single phoneme /j/, or one or both sounds may be categorized as foreign, non-English phones. This dissertation did not test English speakers with a categorization task to determine exactly how participants were categorizing the specific tokens used in these experiments, but ideally follow-up work should be conducted to provide a more complete picture of the differences between the stimuli used in Experiments A1-A3. Additionally, follow-up discrimination tasks should also be conducted so that the perceptual distinctness of critical stimuli used in these experiments can be quantitatively compared.

## 7. Discussion of Experiments A1-A3: Bias and Sensitivity

Experiments A1-A3 show that distributional learning *can*, but does not *necessarily*, affect learners' sensitivities to stimuli.

### 7.1. CURRENT PROPOSAL: A TWO-STAGE MODEL

This section focuses on the results of Experiments A1-A3 to put forth a model of distributional learning.

In Section 2.3, two hypotheses were outlined regarding the driving mechanism behind distributional learning: the Sensitivity Hypothesis, and the Bias Hypothesis, illustrated in Figure 28. Models which base distributional learning in perceptual warping, such as Guenther and Gjaja (1996) and Boersma et al. (2003), assume a sensitivity-based account. In these models, each token experienced by a language learner serves to change the perceptual system, either in synapse weights (as in Guenther and Gjaja, 1996) or in constraint rankings (as in Boersma et al., 2003). This is seen in the top image in Figure 28. In a bias-based account, learners form rough hypotheses regarding the number of categories in the speech stream before identifying category boundaries or category means.

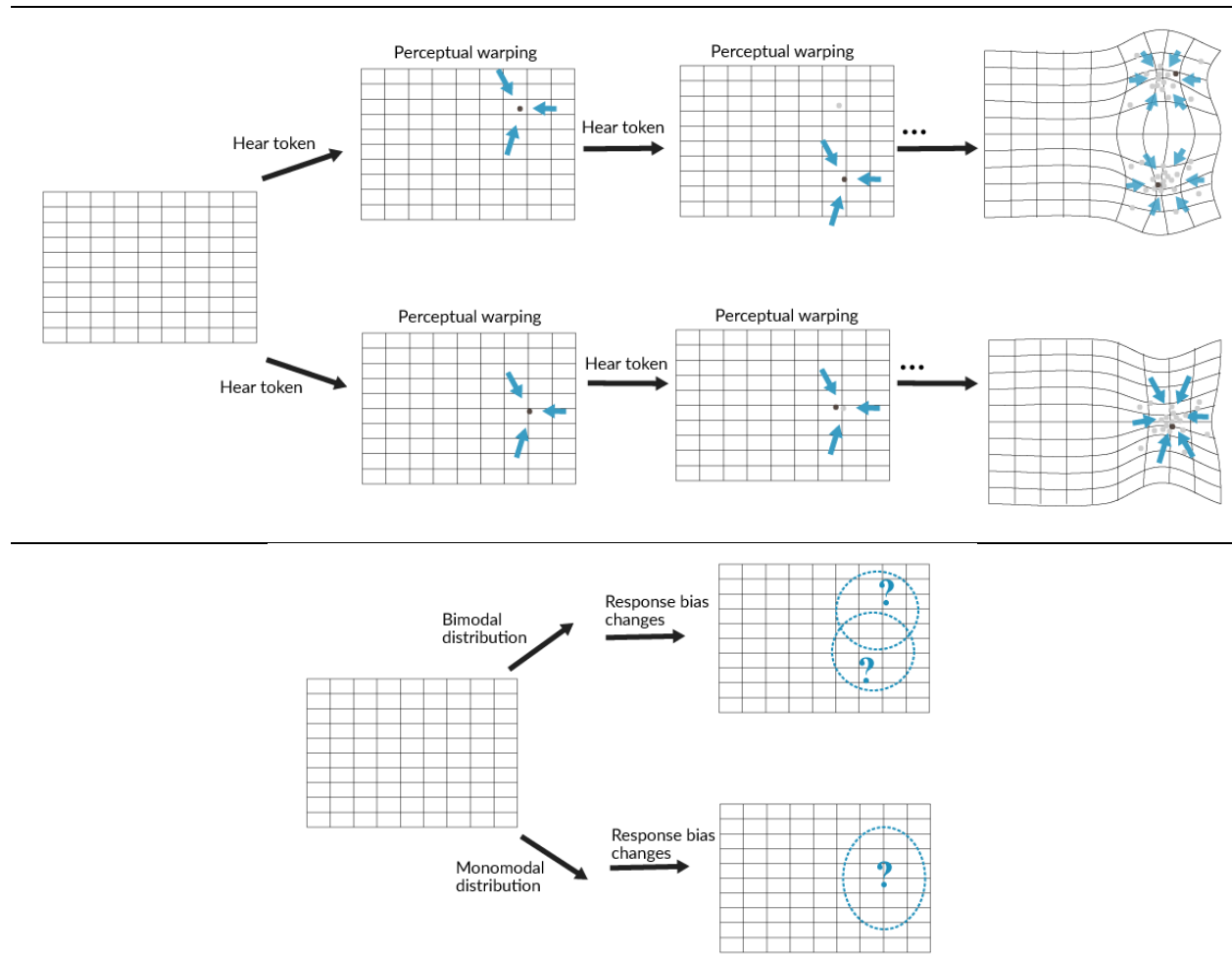


Figure 28. Illustration of the Sensitivity Hypothesis (top) and Bias Hypothesis (bottom).

The set of experiments conducted in this chapter finds support for sensitivity changes, as well as changes in response bias. Specifically, Experiment A3 finds support for a change in sensitivity, and Experiments A1 and A2 find support for changes in bias. The following proposals are considered:

- 1) **Hypothesis 1: Learning method differs for different stimuli.** Learners make use of at least two methods when acquiring phonetic categories: sensitivity-based distributional learning, and bias-based distributional learning. The method used depends on properties of the stimuli.
- 2) **Hypothesis 2: A change in sensitivity precedes a change in bias.** Distributional learning occurs in stages, with a change in sensitivity occurring first, and a change in bias occurring

second. The speed at which learners proceed through each stage depends on properties of the stimuli.

- 3) **Hypothesis 3: A change in bias precedes a change in sensitivity.** Distributional learning occurs in stages, with a change in bias occurring first, and a change in sensitivity occurring second. The speed at which learners proceed through each stage depends on properties of the stimuli.

Although further research is required to test each of these hypotheses, this section will present evidence for Hypothesis 3. This evidence will come in the form of previous studies' results which were reported as non-replications. In doing so, this chapter will also provide support that the results of Experiments A1-A3 may not be limited to artificial language learning studies conducted on adults, and may also be applicable to first language acquisition.

In Chapter 2, it was noted that there was an issue in studying distributional learning with artificial language learning tasks with adults. Namely, although studies like the current one test the effect of distribution on adults, the motivation for distributional learning lies in observations regarding language development in infants – we know that infants cannot be acquiring phonetic categories through minimal pairs because they show language-specific discrimination of phonetic categories by 12 months of age (Kuhl et al., 2006; Mattock and Burnham, 2006; Werker and Tees, 1984; Seidl et al., 2009; Cheour et al., 1998; Polka and Werker, 1994), but only know an estimated 36 words by 8 months of age, none of which are minimal pairs (Caselli et al., 1995), and still confuse minimally-different words like *bih* and *dih* at 14 months (Werker et al., 2002; Stager and Werker, 1997; although see Yoshida et al., 2009; Rost and McMurray, 2009; 2010). Although this chapter acknowledges that the current study is unable to make claims about infant language development for this reason, the results of several past infant studies seem to support this chapter's proposal.

In order to test the effect of distribution on infants, Maye et al. (2002) measure looking times after distributional training. The test phase consisted of both alternating trials and non-alternating trials. In

alternating trials, infants were exposed to a string of tokens alternating between endpoints of the critical continuum they had been exposed to during training ( $D_1a...D_8a...D_1a...D_8a...$ ). In non-alternating trials, infants were exposed to a string of identical tokens, either  $D_3$  ( $D_3a...D_3a...D_3a...$ ) or  $D_6$  ( $D_6a...D_6a...D_6a...$ )<sup>15</sup>. Infant looking times were measured for both TrialTypes, under the assumption that the more attuned an infant is to the difference between the endpoints  $D_1$  and  $D_8$ , the greater the difference in looking times to alternating than non-alternating trials due to a novelty preference for alternating trials. Both 6-month old and 8-month old infants were tested. Maye and colleagues mainly highlight their finding that, when the 6-month and 8-month infants' data are pooled together, there is a significant simple effect of TrialType (alternating or non-alternating) in the Bimodal condition (with Bimodal infants looking longer at alternating trials than non-alternating trials), but not the Monomodal condition. They interpret this as increased sensitivity in the Bimodal condition to the difference between TrialTypes. However, they do also report a significant main effect of Distribution within both the 6-month olds and 8-month olds data, with both age groups having longer looking times on alternating *and* non-alternating trials if trained on a bimodal distribution than if trained on a monomodal distribution ( $p < 0.05$  for both age groups). We can see this in Figure 29, which graphs data reported in Maye et al. (2002). Note that the Bimodal conditions for both age groups have longer mean looking times for both alternating and non-alternating trials.

---

<sup>15</sup> Continuum points 1, 8, 3, and 6 were all experienced the same number of times by the Bimodal and Monomodal conditions (see Figure 15 on page 3).

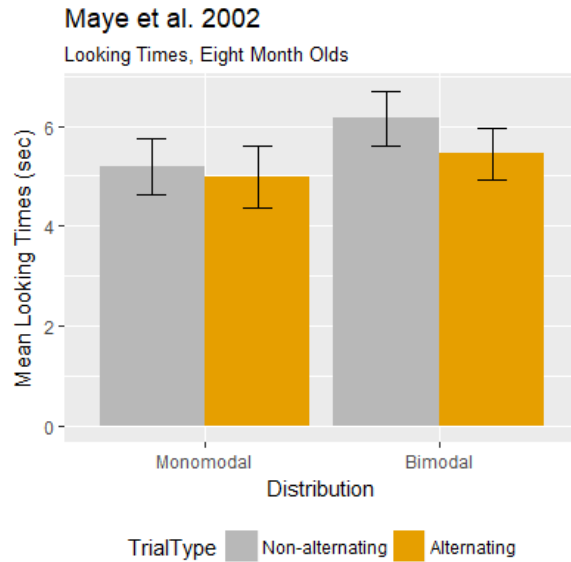
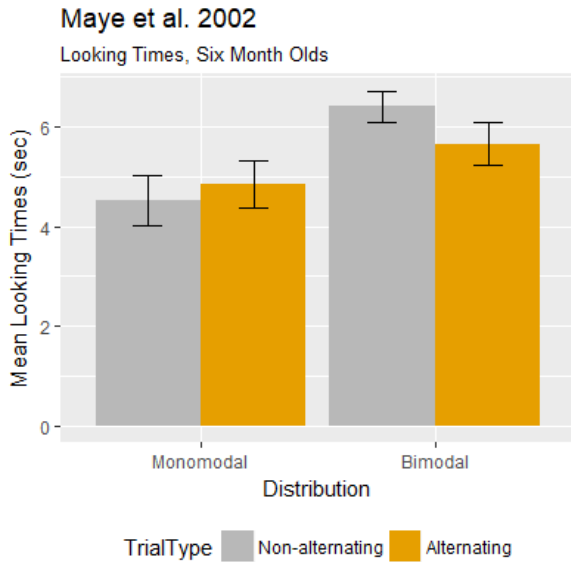


Figure 29. Looking times for infants in Maye et al. (2002).

Similar results are reported in Yoshida et al. (2009), who also measure looking times. Yoshida et al. conduct three experiments with 10-month old infants. Only Experiment 1 and Experiment 3, which use the same stimuli from Maye et al. (2002), will be discussed here.<sup>16</sup>

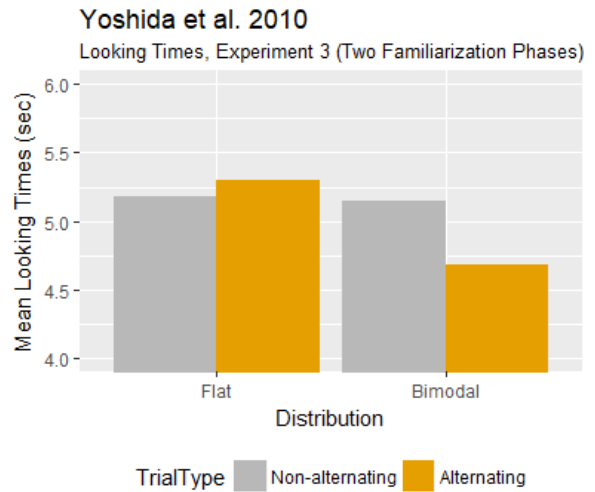
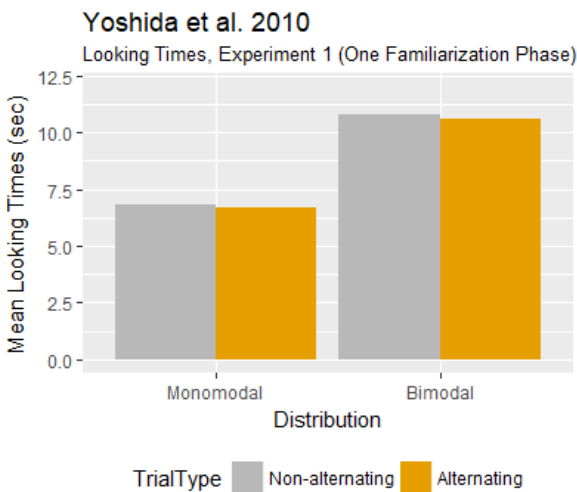


Figure 30. Looking times for infants in Experiments 1 and 3 in Yoshida et al. (2010).

<sup>16</sup> Experiment 2 tests a different contrast (retroflex-dental).

Yoshida and colleagues note a significant main effect of Distribution in Experiment 1 ( $p = 0.002$ ), with the Bimodal group looking longer than the Monomodal group at both alternating and non-alternating trials. Despite this main effect, the authors report the results of Experiment 1 as a non-replication of Maye et al. (2002) since the Bimodal group does not show a significant difference between alternating and non-alternating trials. Yoshida and colleagues suggest that 10-month olds are more resistant to distributional learning than the 6-8 month olds tested in Maye et al. (2010) and go on to test the effects of a longer familiarization phase in Experiment 3.<sup>17</sup> In Experiment 3, they find the predicted simple effect of TrialType for the Bimodal condition (0.018), but not for a Non-modal (“Flat”) condition. (see Figure 30).<sup>18</sup>

Yoshida et al.’s study provides us with evidence that a change in bias and a change in sensitivity occur **sequentially**. Infants trained for just one familiarization phase exhibit a difference in bias, while infants trained for two familiarization phases exhibit a difference in sensitivity. Although one could claim that the sensitivity finding in Experiment A3 was a result of the use of a fricative contrast rather than the stop contrast used in Experiments A1 and A2, the results from Yoshida et al. (2010) lend support to the two-stage model of distributional learning described in Hypothesis 3: learners’ biases change first, as seen in Yoshida and colleagues’ Experiment 1, and we only find a change in sensitivities after longer familiarization times, as seen in Yoshida and colleagues’ Experiment 3.

Returning to all three hypotheses, I would like to note that although Yoshida et al.’s study lends support to Hypothesis 3, this support should not disqualify Hypothesis 1 from also being true. There are many differences between the critical stimuli used in Experiments A1/A2 and those used in Experiment A3, and it is possible that one of these differences or a combination of these differences led participants to use different learning methods. Below I will point out some stimulus differences that may have resulted in

---

<sup>17</sup> The third experiment also differs from the first experiment in that infants are exposed to either a bimodal distribution or a flat, non-modal distribution.

<sup>18</sup> Interestingly, they also find a significant main effect using a different contrast in Experiment 2, but find that the Bimodal condition has *shorter* looking times than a Flat condition.



the use of different learning methods. However, it should be noted that this is speculation, since I did not carry out any direct comparison studies using my specific stimuli and any studies referenced below reached conclusions based on their own set of stimuli which of course differ from my own. To summarize, I consider the following differences as having a possible effect on learning method: 1) a difference in degree of categorical perception between stops and fricatives, 2) a difference in the level of naturalness between my stop and fricative continua, and 3) a (likely) difference in perceptual distance between continuum endpoints.

There are a number of differences between the critical stimuli used in Experiments A1/A2 and those used in Experiments A3, most notably that the critical stimuli in Experiments A1 and A2 were stops, whereas the critical stimuli used in Experiment A3 were fricatives. It is possible that fricatives by nature are perceived less categorically than stops are. Both stops and fricatives appear to be perceived categorically at least in some cases (Strand and Johnson, 1996; Repp, 1981; Sharma and Dorman, 1999), but it is possible that the degree of categorical perception differs between different pairs of speech sound. No study I know of directly compares the degree of categorical perception between stops and fricatives though, and a comparison would need to be completed with my specific stimuli to see if this may have contributed to the different findings in Experiments A1/A2 and Experiment A3. Along these lines, Repp (1981) suggests that fricatives are perceived categorically when perceived as speech, but gradiently when perceived as non-speech noise. If this observation is special to fricatives, it's possible that some participants perceived the fricative stimuli as non-speech noise, and therefore in a gradient manner. Again, a direct comparison of the perceptual properties of my specific stimuli would need to be completed to determine if there is a significant difference in perception of the stop and fricative critical stimuli used in these experiments. The idea of different modes of processing for different classes of phones is not a new one. Toro and colleagues (2008) argue for two different modes of processing of vowels and consonants in what they refer to as the "CV Hypothesis." They support their hypothesis by showing that participants will generalize template-like rules for vowels (i.e. generalizing the ABA pattern in the word *tapena* to

*biduki*), but not for consonants. Consonants, on the other hand, are used to extract words from a speech stream. If the stops were perceived more categorically than the fricatives were, the fricative stimuli may have been more conducive to a shift in sensitivity, which may be why Experiment A3 is the only experiment we find any evidence for sensitivity changes.

It is also possible that factors unrelated to phone type played some role. Blomert and Mitterer (2004) find that speech continua which sound more natural are perceived less categorically than continua which sound synthetic. While the critical continua I used sounded similar in terms of naturalness to my ears, it could have been the case that one was more natural than the other, leading to different degrees of categorical perception between critical continua.

It is also likely that the perceptual distance between continuum endpoints differed between experimental stimuli. I did not carry out any discrimination tests to measure Just-Noticeable Differences (JNDs) between endpoints, but based on my own listening of these stimuli, I would predict that the endpoints of the fricative stimuli are more perceptually distinct than the endpoints of the stop stimuli, which conceivably could have resulted in participants utilizing different learning methods for different critical continua.

This chapter proposes a two-stage model for the acquisition of phonetic categories. In the first stage, learners form a rough idea of the number of phonetic categories in the speech stream. During this stage, the distribution of phones encountered affects learners' **response bias**, such that exposure to a bimodal distribution of phones causes learners to expect two phonetic categories and a monomodal distribution of phones causes learners to expect one phonetic category. This is illustrated in Figure 31.

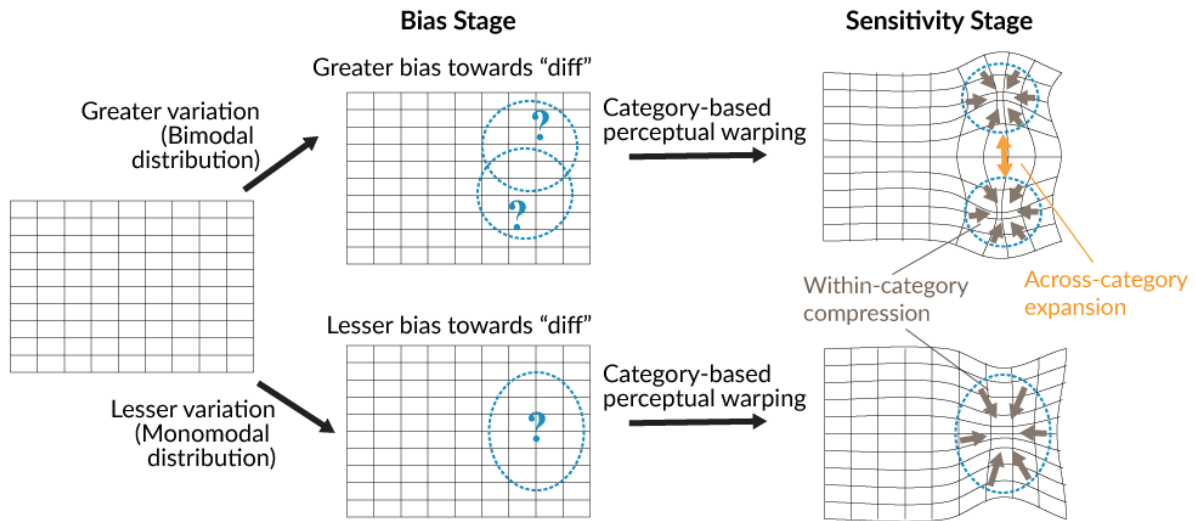


Figure 31. Illustration of proposed mechanism behind phonetic category acquisition.

In this Bias Stage, the learner only has a general idea of the number of phonetic categories, and does not have well-defined category means or category boundaries. A greater expectation or assumption that two categories exist results in more overall responses that a given pair of critical stimuli are “different,” whether or not they are actually different or not. This can be seen in Experiments A1 and A2, where bimodally-trained participants were more likely to respond “different” compared to monomodally-trained participants, even for pairs which were identical (e.g. G<sub>1a</sub> vs. G<sub>1a</sub>).

What is responsible for the difference in these expectations or assumptions held by the bimodally- and monomodally-trained language learners? Although further research is required to determine the root cause of these differing assumptions, it is suggested here that this can be explained by how much participants notice variation in the training phase. Since the bimodal group hears more tokens at the continuum points 2 and 7 and the monomodal group hears more tokens at continuum points 4 and 5, it is more likely that a participant during bimodal training will encounter continuum point 2 followed by continuum point 7 (or vice versa) than a participant during monomodal training, who is more likely to encounter continuum point 4 followed by continuum point 5. Because continuum points 2 and 7 are more acoustically distinct, a bimodally-trained participant may be more consciously aware of the difference between critical

stimuli than a monomodally-trained participant, simply because they hear two distinct tokens presented side-by-side more often than the monomodal participants. Even if the bimodal participants are not able to identify what the acoustic difference between these two sounds might be, we could imagine they might have a stronger awareness that there at least *are* two different sounds compared to the monomodal participants, resulting in more “different” responses when the bimodal participants are presented with pairs of critical syllables. There are at least two ways this could be tested. The first would be to train participants on a bimodal or monomodal distribution along one phonetic dimension, but then test participants with pairs of sounds which differ along a different phonetic dimension. If participants with a greater bias towards a “different” response believe there are two different sounds but truly do not know how they differ, participants would be expected to respond that pairs of sounds are different even along an untrained dimension.<sup>19</sup>

A second way to test this would be to present participants with non-random bimodal distribution of phones. Specifically, one group of participants could be exposed to a bimodal distribution in which the continuum points 2 and 7 occur side by side very often (e.g. in the order 2-7-2-7-1...etc.). Another group of participants could also be exposed to a bimodal distribution of phones, but in an order such that only consecutive continuum points are presented side by side (e.g. in the order 1-2-2-2-3-4-5-6-7-7-7-8...etc.)

Supposing the differences in bias can be explained as an awareness of variation in the speech signal, one would still need to explain how some distributional learning experiments result in greater sensitivity in bimodally-trained participants compared to monomodally-trained participants. This second stage could potentially be explained with some perceptual warping distributional learning mechanism such as those suggested by Guenther and Gjaja (1996) or Boersma et al. (2003), but I believe this can be explained with domain-general mechanisms which have already been thoroughly documented in the psychology literature: **across-category expansion** and **within-category compression** (Livingston et al.,

---

<sup>19</sup> Many thanks to Stefan Gries for this interesting suggestion for future research.

1998; Goldstone, 1994; Goldstone and Hendrickson, 2010; Nosofsky, 1986). That is, the formation of some category warps perception such that items within a category are perceived as being more similar to one another, and items from different categories are perceived as being more dissimilar from one another. This has been found for visual stimuli, such as drawings of hypothetical microorganisms (Livingston et al., 1998), drawings of chick genitalia (Livingston et al., 1998), faces of well-known (Beale and Keil, 1995) and unknown faces (Levin and Beale, 2000), drawings of rock formations (Kurtz and Gentner, 1998), squares varying in brightness and size (Goldstone, 1994), shapes (de Beeck et al., 2003), and colors (Winawer et al., 2007; Özgen and Davies, 2002). Infants exposed to individual monkey faces each with a different label show a novelty preference when shown an unseen monkey face, whereas infants exposed to those same monkey faces either without a label or all labeled as “monkey” do not (Scott and Monesson, 2009). This has also been found for (non-linguistic) auditory stimuli such as white noise samples (Guenther et al., 1999) and musical chords (Burns and Ward, 1978).

To summarize, this section proposes that learners exposed to a bimodal distribution of phones are presented with more examples of stimuli which are different enough to be noticeable. This leads to an increase in awareness that there are two different sounds in a **Bias Stage** of distributional learning. Learners who are more aware that there are two different sounds are more likely to attempt to categorize these sounds into different categories, and category-based perceptual warping results in greater sensitivity to sounds which cross category boundaries in a **Sensitivity Stage**. Neither of these stages require the proposal of a special mechanism for phonetic category acquisition.

## 8. “Tone” Experiments

As mentioned earlier, the original purpose of the “A” Experiments was to simply attempt to replicate Maye and Gerken (2000) over the web. One pilot experiment, not reported here, was initially conducted with this goal in mind. This pilot experiment differed from Maye and Gerken (2000)’s study in a number of ways. In particular, it was thought that a number of catch trials should be included throughout the experiment in order to ensure that participants taking this experiment in some unknown setting would be

paying attention to stimuli. Although this pilot differed from Experiments A1-A3 in a number of ways, one of these differences seemed worth following up on. That is, the pilot experiment failed to show any significant differences in responses between bimodally- and monomodally-trained participants during the Test phase. It was hypothesized that one methodological difference in particular might be responsible for this: during the Train phase, catch trial tones were randomly interspersed with stimuli. This concurrent Train Catch task had the goal of ensuring that participants were wearing headphones and paying attention. To do this, each Train repetition contained 6 randomly-interspersed catches: 3 one-tone tokens and 3 two-tone tokens. Participants were instructed to press the “1” or “2” keys if they heard one of these tone tokens, to indicate how many tones they had heard. Tones were chosen to be at a low enough frequency that most computer speakers would not pick up on the sound (50 Hz), thereby testing whether participants were wearing headphones or not. Tones were 340 ms long, and were 140 ms apart for the two-tone tokens. Participants were given the following instructions:

*For the most part, you will be listening passively and will not need to click on anything. However, to help you keep your attention on the task, you will hear one or two low-toned beeps randomly-interspersed throughout. As quickly as possible, please press “1” if you hear one beep, and “2” if you hear two beeps.*

In order to follow up on whether or not the inclusion of this monitoring task had an effect on distributional learning, two further experiments were designed which were identical to Experiments A2 and A3, with the exception that each included these Train Catch tones during the Train phase in addition to the original Train stimuli. These two experiments are called Experiments A2-Tone and A3-Tone respectively.

The following exclusion criteria were used:

- Fewer than 5/6 on the pre-experiment Sound Check (Exp A2-Tone: 14 excl, Exp A3-Tone: 21 excl<sup>20</sup>)
- Fewer than 5/8 correct on the Practice Test (Exp A2-Tone: 0 excl, Exp A3-Tone: 8 excl)
- Reported not being a native speaker of English (Exp A2-Tone: 0 excl, Exp A3-Tone: 2 excl)
- Reported a history of a speech or hearing disorder (Exp A2-Tone: 0 excl, Exp A3-Tone: 5 excl)

---

<sup>20</sup> A large number of participants were excluded for this reason. See the proposed explanation in Footnote 13.

In addition, 5 participants in Experiment A3-Tone were excluded from analysis for reporting having experience with a language with more than one voiceless post-alveolar fricative during the questionnaire. In total, 14 participants were rejected from analysis from Experiment A2-Tone, leaving 28 in the Bimodal group and 31 in the Monomodal group. 28 participants were rejected from analysis from Experiment A3-Tone, leaving 24 in the Bimodal group and 19 in the Monomodal group.

## 8.1. METHODS

The procedure and stimuli of Experiments A2-Tone and A3-Tone were identical to their Experiments A2 and A3 respectively, with the exception of the inclusion of a concurrent Train Check “beep”-monitoring task in these Tone Experiments. Each Train repetition consisted of 3 one-tone tokens and 3 2-tone tokens. Each Train repetition was repeated 4 times, resulting in a total of 192 fillers, 192 critical tokens, and 24 beep tokens.

## 8.2. RESULTS: EXPERIMENT A2-TONE

A generalized linear mixed models with a logit link function (GLMM) was fitted to the formula in (5), with the reference cell being *bimodal diff*. For Experiment A2-Tone, the model using the default Laplace Approximation algorithm failed to converge in 10,000 evaluations. Because of the failure to converge, an Adaptive Gauss-Hermite Quadrature algorithm was used instead (by setting `nAGQ` to 0 in R). Summaries of the fixed effects in the mixed logit model with treatment coding for Experiment A2-Tone are shown in Table 13. For Experiment A2-Tone, no interaction between *Distribution* and *PairType* was found for critical trials ( $p = 0.551$ ), or for filler trials ( $p = 0.986$ ).

Predictor	Coefficient	SE	Wald Z	<i>p</i>
<b>CRITICAL</b>				
(Intercept)	-2.170	0.354	-6.137	<0.001 ***
Distribution= <i>monomodal</i>	-0.189	0.483	-0.391	0.696
PairType= <i>same</i>	-1.584	0.426	-3.720	<0.001 ***
Interaction= <i>monomodal &amp; same</i>	-0.197	0.598	-0.329	0.742
<b>FILLER</b>				
(Intercept)	3.736	0.557	6.714	<0.001 ***
Distribution= <i>monomodal</i>	0.101	0.800	0.127	0.899
PairType= <i>same</i>	-6.551	0.715	-9.161	<0.001 ***
Interaction= <i>monomodal &amp; same</i>	-0.217	1.034	-0.210	0.834

Table 13. Summary of fixed effects in the mixed logit model in Experiment A2-Tone.



Figure 32. Log odds of participants responding *d* for critical trials (left) and filler trials (right) in Experiment A2-Tone. Error bars indicate standard error.

To test for a main effect of Distribution on Response, a planned contrast analysis was performed on the fitted GLMM using the `glht` function from the `multcomp` package (Hothorn et al., 2017) in R. The planned contrast revealed no main effect of Distribution for critical trials ( $p = 0.551$ ) or filler trials ( $p = 0.986$ ).



### 8.3. RESULTS: EXPERIMENT A3-TONE

A generalized linear mixed models with a logit link function (GLMM) was fitted to the formula in (5), with the reference cell being *bimodal diff*. For Experiment A3-1, the model using the default Laplace Approximation algorithm failed to converge in 10,000 evaluations. Because of the failure to converge, an Adaptive Gauss-Hermite Quadrature algorithm was used instead. Summaries of the fixed effects in the mixed logit model with treatment coding for Experiment A3 are shown in Table 14. For Experiment A3, no interaction between Distribution and PairType was found for critical trials ( $p = 0.742$ ), or for filler trials ( $p = 0.834$ ). Results are shown in Figure 33.

Predictor	Coefficient	SE	Wald Z	$p$
<b>CRITICAL TRIALS</b>				
(Intercept)	1.027	0.684	1.500	0.134
Distribution= <i>monomodal</i>	-0.822	1.032	-0.797	0.425
PairType= <i>same</i>	-4.349	0.703	-6.185	<0.001 ***
Interaction= <i>monomodal &amp; same</i>	-0.457	1.173	-0.390	0.697
<b>FILLER TRIALS</b>				
(Intercept)	2.799	0.487	5.753	<0.001 ***
Distribution= <i>monomodal</i>	1.433	0.982	1.459	0.145
PairType= <i>same</i>	-4.837	0.593	-8.159	<0.001 ***
Interaction= <i>monomodal &amp; same</i>	-1.535	1.121	-1.370	0.171

Table 14. Summary of fixed effects in the mixed logit model in Experiment A3-Tone.

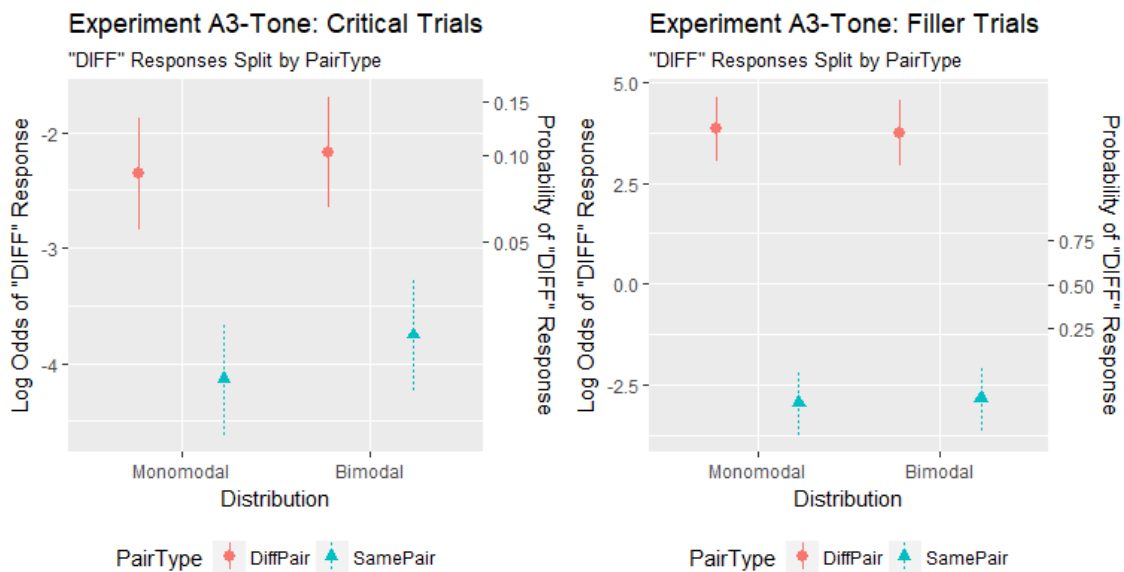


Figure 33. Log odds of participants responding *d* for critical trials (left) and filler trials (right) in Experiment A3-Tone. Error bars indicate standard error.

To test for a main effect of Distribution on Response, a planned contrast analysis was performed on the fitted GLMM using the `glht` function from the `multcomp` package (Hothorn et al., 2017) in R. The planned contrast revealed no main effect of Distribution in Experiment A3-Tone for critical trials ( $p = 0.151$ ) or filler trials ( $p = 0.222$ ).

#### 8.4. DISCUSSION

The addition of Train Catch tones during the Train phase appeared to have a negating effect on distributional learning as Experiments A2-Tone and A3-Tone, which only differed from Experiments A2 and A3 in the inclusion of the beep tokens, failed to show a significant effect of Distribution.

### 9. Discussion: Attention in Distributional Learning

For each of the tone/no-tone counterpart experiments conducted here, experiments which contained Train Catch tones resulted in no significant difference between bimodally-trained and monomodally-trained participant responses to critical stimuli. This seems to suggest that attention plays some sort of role in distributional learning. Although this study is not equipped to say what this role may be, this section provides support for two competing hypotheses: 1) that attention facilitates distributional learning, and 2) that attention impedes distributional learning.

Before providing support for these two hypotheses, this section briefly reports on results from the after-experiment questionnaire to determine how participants self-reported their attention levels. The questionnaire in all experiments asked participants how much attention they were paying to various phases of the experiment, and whether or not they would pay more, the same, or less attention if this same experiment were being conducted in a lab setting. Participants were given the following options, for both the Train phase and the Test phase: (a) *I focused all of my attention on this portion of the experiment*, (b) *I mostly paid attention*, (c) *I was not paying very much attention*, or (d) *I paid very little attention*. Very few participants reported “not paying very much attention” or reported paying “very little attention.” A

glance at how participants responded to options (a) and (b) suggests that the inclusion of Catch trials did not have a large impact on reported attention levels. A breakdown of responses is shown in Table 15.

		Did not contain Train Catch tones			Contained Train Catch tones	
		Exp. A1	Exp. A2	Exp. A3	Exp. A2-Tone	Exp. A3-Tone
Train phase	I focused all of my attention on this portion of the experiment	58%	52%	78%	68%	61%
	I mostly paid attention	38%	47%	20%	30%	37%
Test phase	I focused all of my attention on this portion of the experiment	84%	85%	88%	95%	84%
	I mostly paid attention	15%	15%	12%	5%	16%

Table 15. Questionnaire responses regarding participants' attention.

Although statistical tests were not performed, numerically participants in all experiments reported paying more attention during the more active Test phase, with a greater percentage of participants reporting focusing “all” of their attention to that portion (specifically, 84-95%). No clear difference in self-reported responses between Tone and No Tone experiments can be seen. Speculation regarding the inclusion of tones and their role in participant responses is given below.

### 9.1.1 *A case for attention facilitating distributional learning*

Although catch tones were included to maintain participants' attention on the task at hand, the addition of the tone-monitoring task may have *decreased* participants' attention to the syllables they were being exposed to by taking away attentional resources from the training phase. If so, this lack of attention to the training data may have been responsible for the absence of distributional learning for Tone experiments. In other words, this training task may have in essence become a task in which participants were listening for tones and treating actual stimuli as outside noise to be filtered. A similar finding has been made for speech segmentation. Learners are able to make use of transitional probabilities to segment a stream of speech (Saffran et al. 1996), but if their attention is diverted, they exhibit less learning. Toro et al. (2005) and Saffran et al. (1997) both exposed learners to a speech stream in a segmentation experiment. Both

found that learners were able to successfully make use of transitional probabilities if attention was diverted to a task with little demand that did not make use of the same sensory modality (like drawing while listening to the speech stream). However, Toro et al. found that more demanding tasks or tasks that made use of the same sensory modality (that is, a concurrent auditory task) negatively affected participants' abilities to segment speech using transitional probabilities.

### 9.1.2 *A case for attention impeding distributional learning*

On the other hand, it could be argued that the tone-monitoring task increased participants' attention to the training phase, and that this increased attention impeded distributional learning in the Tone experiments. Cutler et al. (1987) make a distinction between comprehension-oriented attention and perception-oriented attention. Some researchers find evidence that increased perception-oriented attention prevents listeners from shifting phonetic category boundaries to allow for talker variation (McAuliffe and Babel, 2016; Pitt and Szostak, 2012). For example, Pitt and Szostak (2012) played English words to participants, replacing [s] with [ʃ] and [ʃ] with [s]. In a following lexical decision task, participants who were told to simply listen during the training phase were more tolerant of variation, and were more likely to respond that non-words such as [s]andelier and [ʃ]erenade were English words, compared to participants who had been explicitly told that they should pay attention to the pronunciation of the speaker before the training phase. Pitt and Szostak conclude that participants were more able to shift their phoneme category boundaries if their (perception-oriented) attention had not been directed to the talker's pronunciation.

Again, this set of experiments is not equipped to say whether attention facilitated or impeded distributional learning. Interestingly though, Ong et al. (2015) appears to find results which contradict those presented here. In a distributional learning experiment using Thai tones, Ong and colleagues fail to find evidence for distributional learning *unless* they had a concurrent task in which they monitored non-linguistic acoustic tones (or "beeps"). These divergent results may be due to the differing natures of the critical stimuli used in each experiment (perhaps an acoustic monitoring task has a different effect on supersegmentals such as lexical tone), or may be due to the differing natures of the monitoring stimuli

themselves. The monitoring stimuli used here were very low-frequency pure tones, and so may have been more difficult to detect than the monitoring stimuli used by Ong and colleagues. Further research is needed to determine what role the monitoring task played in these experiments.

## **10. Discussion: Distributional Learning Online**

In addition to the above theoretical contributions of this chapter, this study also makes two methodological suggestions for those wishing to conduct phonetic experiments on Mechanical Turk. This section concludes that (1) it is possible to replicate results of studies that require fine phonetic distinctions on MTurk, and (2) changes made to adapt an experiment from a lab-based experiment to an online experiment should be kept to a minimum.

Regarding (1), I believe that the inclusion of a short task confined to the beginning of an MTurk experiment, particularly a task which requires participants to listen for sounds which most computer speakers cannot pick up (50 Hz non-linguistic tones in this case), is sufficient encouragement to participants to wear headphones for the duration of the experiment. Questions were included in the post-experiment questionnaire for all “A” Experiments to determine whether participants were actually wearing headphones. Participants were specifically told that their answers would not affect their payment. Participants were asked to respond whether they were (a) wearing headphones the entire time, (b) wearing headphones most of the time, (c) wearing headphones some of the time, or (d) not wearing headphones at all. Most of the participants reported that they were wearing headphones the entire time, with only a few reporting wearing headphones only “most of the time,” (one each in Experiments A1, A2, and A2-Tone, and three in A3-Tone), and only one participant reported wearing headphones “some of the time” or not wearing headphones at all (in Experiment A3).

In attempting to replicate distributional learning online, it was initially believed that certain changes needed to be made to ensure that participants were paying attention and wearing headphones. However, results of this study suggest that forcing participants to pay attention to non-linguistic intervening stimuli may have had unintended effects, at least for distributional learning. Therefore, it is suggested

that only minimal changes, such as including a short sound check at the beginning of the experiment, should be made when attempting to replicate studies over the web.

## **11. Discussion: Filler Trials**

This study found some unexpected results with regards to filler trials. Since filler stimuli were identical across conditions, it was not expected that there would be any significant differences in participant responses for Test filler pairs. However, the following two unexpected results were found:

- 1) A significant interaction between Distribution and PairType was found for filler trials in Experiment A1 (greater sensitivity to fillers in the monomodal condition).
- 2) A significant interaction between Distribution and PairType was found for fillers in Experiment A3 (greater sensitivity to fillers in the bimodal condition).

Although further research is needed, this section provides some speculation as to why we see an effect on filler trials.

First, it should be noted that the only two instances where filler trials exhibited any significant results were in experiments for which the author created the stimuli. That is, Experiment A1 and A3 used stimuli created by the author, but Experiment A2 used stimuli from Maye and Gerken (2001). I believe the main difference in these stimuli had to do with how much the repetitions of “same” stimuli varied. These experiments followed up on the methodology of Maye and Gerken (2000) who asked participants whether a pair of stimuli were repetitions of the “same” syllable or were “different” syllables. Because of this, it was decided beforehand that these experiments should not be discrimination tasks, in which participants would respond “different” if they could detect any difference whatsoever in the acoustics of a given pair of stimuli. Therefore when creating stimuli, I purposely made sure that repetitions of filler syllables (e.g. [fa]<sub>1</sub> vs. [fa]<sub>2</sub>) were distinguishable as different tokens. However, upon listening later to the stimuli provided by Maye and Gerken, it was determined that the filler tokens used by Maye and Gerken were not different enough to be distinguished by a participant as being two different tokens of the same syllable. Because filler Same Pairs were so similar as to be perceptually identical, it would be unlikely that any

participant would respond that a filler Same Pair was actually “different.” Because of this, I believe the filler stimuli used by Maye and Gerken exhibit a floor effect. However, this only explains why my filler stimuli sometimes behaved differently from those used by Maye and Gerken. Why would training change how participants responded to fillers?

Although further research is needed, I believe the significant differences across conditions for filler trials come from increased attention to small variations depending on which condition a participant is in. If it is the case that bimodally-trained participants pay more attention to stimuli because of the greater number of 2-7 or 7-2 pairs during training, it is possible that this increased attention bled over into the filler stimuli. This increased attention to variation may have caused participants to form various hypotheses regarding the filler stimuli. As for why the fillers in Experiment A1 and Experiment A3 show opposite behaviors, that may simply be because the fillers across these two experiments were different. How participants treated them due to increased attention to variation may come from a number of factors, such as what dimension(s) filler tokens varied from one another, how noticeable the variation between filler tokens was, and how similar or dissimilar filler tokens were from critical tokens.

## 12. Conclusion

In a set of five artificial language learning experiments conducted through Mechanical Turk, this chapter makes a distinction between **bias** and **sensitivity**. A brief summary of the procedure and results of the “A” Experiments<sup>21</sup> can be found in Table 16.

---

<sup>21</sup> Two other experiments were conducted, but the results and methodology of these experiments are not included in this dissertation. One of these experiments tested all participants over the age of 18 who volunteered (rather than being restricted to the ages of 18-25), and another experiment used a vowel continuum, rather than a stop continuum, as its critical stimuli. Details of these experiments can be found in Moeng (2017). The experiment which tested all participants over the age of 18 found an “anti” distributional learning effect, with bimodal participants responding *less* often than monomodal participants that critical stimuli were “different” from one another. Details of this experiment can be found in Moeng (2017).

	A1 Experiment	A2 Experiments		A3 Experiments	
	Exp. A1	Exp. A2	Exp. A2-Tone	Exp. A3	Exp. A3-Tone
Stimuli	Created by author	Originally used by Maye and Gerken (2001)		Created by author	
	G1a-G8a G1ae-G8ae G1r-G8r			S1a-S8a	
Procedure	Sound Check				
	Practice Test				
	Train phase	Train phase	Train phase + catch trials	Train phase	Train phase + catch trials
	Test phase	Test phase	Test phase	Test phase	Test phase
	Questionnaire				
Results: Evidence for distributional learning?	Yes, specifically in <b>response bias</b>	Yes, specifically in <b>response bias</b>	No	Yes, specifically in <b>sensitivity</b>	No

Table 16. Summary of stimuli, procedures, and results for all “A” Experiments.

This study finds evidence for distributional learning through an online platform, in response bias and in sensitivity. Experiments A1 and A2 find that the bimodal group has a greater bias towards a “different” response than the monomodal group, while Experiment A3 finds that the bimodal group has a higher sensitivity than the monomodal group. Additionally, no evidence for distributional learning in either bias or in sensitivity was found for those experiments which contained Train Catch tones, Experiments A2-Tone and A3-Tone.

Additionally, two suggestions were made in Section 7 regarding follow-up experiments to test for evidence of the Bias Stage of this model: 1) an experiment which trains learners on a contrast varying along one phonetic dimension but tests learners on a different phonetic dimension, and 2) an experiment in which participants receive bimodal training such that tokens are arranged in a non-random fashion (either with many 2-7 tokens played side-by-side, or with only consecutive continuum points played side-by-side).



The main contribution of this chapter was to propose a model in which phonetic category acquisition occurs in two stages: first, learners form rough expectations regarding the number of sound categories in the speech stream in an Bias Stage. Following this, learners' sensitivities change through general cognitive mechanisms which drive within-category compression and across-category expansion (Goldstone, 1994) in a Sensitivity Stage. This two-stage model of phonetic category acquisition (not to be confused with the one-stage model of allophony acquisition presented in the next chapter) counters predictions made by previous accounts of distributional learning (Guenther and Gjaja, 1996; Boersma et al., 2003), which predict that a change in sensitivity always accompanies distributional learning. The following chapter will examine the role that contextual environment plays in distributional learning.

## Chapter 4:

### Distributional Learning and Allophony: A One Stage Model of Allophony Acquisition

#### 1. Introduction

Although the previous chapter focused on the acquisition of phonetic categories, the current chapter explores the acquisition of allophonic relationships. Following Dillon et al. (2013), this chapter addresses a model followed by a number of acquisitionists that phonetic category acquisition is the first step in a two-step model of phonological acquisition (for example, see Peperkamp et al., 2003; Peperkamp et al., 2006; Noguchi, 2016; Harris, 1963), as opposed to an alternative model, in which allophones and rules relating allophones to one another are acquired simultaneously (Dillon et al., 2013). In order to address the question of whether phonetic categories and allophonic relationships between those categories are acquired in one stage or in two stages, Experiment B maps the learning trajectory of learners exposed to one of three types of distributions, where learners are trained for either 5 minutes, 10 minutes, or 15 minutes. The following two conclusions are made in this chapter: 1) results from Noguchi (2016) which explores the learning of allophonic relationships can be replicated on Mechanical Turk; and 2) non-significant trends found in this experiment support a one-stage model of phoneme acquisition. Although further research is necessary, this chapter also discusses possible further support for the model proposed in the previous chapter; specifically the two-stage model consisting of a Bias Stage followed by a Sensitivity Stage. The results of experiment in this chapter suggest that participants exposed to a monomodal distribution experience prolonged uncertainty regarding the number of categories in the language, compared to those exposed to a bimodal distribution<sup>22</sup>.

---

<sup>22</sup> Specifically, a “Bimodal-NonComp” distribution rather than a “Bimodal-Comp” distribution, as described later.

## 2. Background

This section will provide background literature on learning allophonic relationships in Section 2.1, and background on past models of allophony acquisition in Section 2.2.

### 2.1. LEARNING COMPLEMENTARY DISTRIBUTION

A number of studies have found evidence that suggests that pairs of sounds which belong to the same phoneme (sound pairs in a “phonemic” relationship) are processed differently than pairs of sounds which belong to the same phoneme (sound pairs in an “allophonic” relationship). In particular, listeners exhibit less sensitivity to sound pairs which are allophonic in their language compared to sounds which are phonemic (Peperkamp et al., 2003; Boomershine et al., 2008; Seidl and Cristia, 2012; Johnson and Babel, 2010). This could be attributed at least in part to the tendency for allophonic sound pairs to exhibit greater perceptual similarity than phonemic pairs (Pegg and Werker, 1997; Yuan and Liberman, 2011), but cannot be solely attributed to perceptual similarity. Boomershine et al. (2008) find that English speaking adults are less sensitive to the difference between pairs of phones which are allophonic in English ([r] and [d]) than they are to pairs of phones which are phonemic in English ([d] and [ð]). However, Spanish speaking adults showed the opposite pattern: they showed low sensitivity to [d] and [ð], which are allophonic in Spanish, and greater sensitivity to [r] and [ð], which are phonemic in Spanish. This suggests that lower sensitivity to allophonic pairs than to phonemic pairs cannot be solely attributed to a lesser perceptual distance between allophonic pairs, since the English and Spanish speakers showed the opposite pattern for the same phones [d] and [ð].

Seidl et al. (2009) finds evidence that the distinction between phonemic and allophonic relationships develops somewhere between 4 and 11 months of age. In an infant language learning study, Seidl and colleagues exposed English-learning and French-learning infants to a phonological pattern which depended on vowel nasality. Crucially, vowel nasality is contrastive in French, but not in English. The English-learning 4-month olds and French-learning 11-month olds were able to learn the pattern, but the

English-learning 11-month olds were not, suggesting that sensitivity to a contrast which is not phonemic decreases by the time an infant is 11 months of age.

The studies above show that phonemic and allophonic pairs are *processed* and *discriminated* differently. However, there are only a few studies which have been designed to catch the acquisition process of an allophonic relationship and its associated change in sensitivity within the lab, the way that Maye and colleagues' distributional learning experiments were able to measure a change in sensitivity caused by different frequency distributions within the lab. The remainder of this section summarizes findings from Peperkamp et al. (2003) and Noguchi (2016), who both test whether the predictability of a phone based on its phonetic environment can bring about a change in participant sensitivity.

Peperkamp et al. (2003) tested three groups of native French speakers: a Monomodal group, a Bimodal group, and a Bimodal+Assimilation group. Critical stimuli consisted of tokens taken from an 8-point continuum ranging between the fricatives [ʁ] and [χ], each preceded by a vowel. These were followed by CV context syllables, which began with either a voiced or voiceless consonant, creating VC<sub>Target</sub> + CV<sub>Context</sub> "phrases." The Monomodal group heard a monomodal distribution of the fricatives [ʁ] and [χ] during the training phase, and both Bimodal groups heard them in a bimodal distribution. The Bimodal+Assimilation group only heard the [ʁ]-half of the continuum before voiced consonants, and the [χ]-half of the continuum before voiceless consonants. During the test phase, participants were presented with pairs of 2-word VC.CV "phrases," and were asked whether the first words in these two phrases were the same or different. This test phase occurred once before the exposure phase, and once after. Peperkamp and colleagues found that the Bimodal group was the only group to show a significant difference between the pre- and post-test phases, but they found no significant interaction across groups. Peperkamp and colleagues suggest that, since the Bimodal group resulted more learning (numerically) between pre- and post-test phases than the Bimodal+Assimilation group, environmental context may play a role in distributional learning. However, given the lack of significant interaction across groups, the authors also caution that the results from their experiment are unclear. They also note that their experiment failed to replicate

Maye and Gerken (2000), since there was no significant interaction between the Bimodal and Monomodal groups. Results may have been affected by the fact that Peperkamp and colleagues tested native speakers of French, who already have the phonological rule specified in the Bimodal+Assimilation group.

In a recent dissertation, Noguchi (2016) tested three groups of participants: a Non-Complementary group, a Complementary group, and a Control group. The first two groups heard a bimodal distribution of critical syllables with the onset ranging from an alveopalatal fricative [ea] to a retroflex fricative [sa] in an 8-point continuum. (The Control group did not hear any of the critical syllables containing fricatives.) The Non-Complementary group heard all 8 points of the continuum following one of four context syllables, all of which ended with [i], and also all 8 points of the continuum following one of four context syllables, all of which ended with [u] (e.g. [liea], [liʂa], [luɛa], and [luʂa]) (see Figure 34, left).

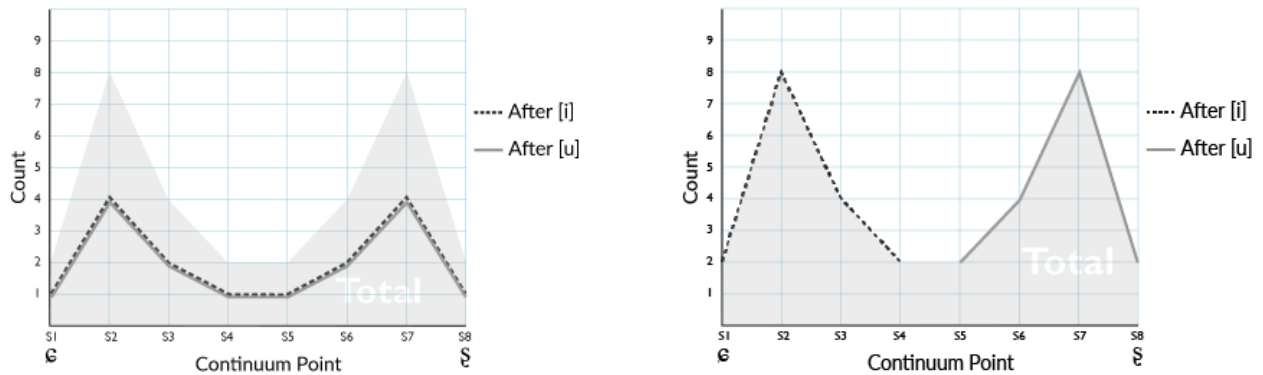


Figure 34. Training distributions for Non-Complementary (left) and Complementary (right) conditions in Noguchi (2016).

The Complementary group only heard the four tokens on the [ea]-side of the 8-point continuum (referred to here as S<sub>1a</sub>-S<sub>4a</sub>) following syllables ending with [i], and the four tokens on the [sa]-side of the 8-point continuum (referred to here as S<sub>5a</sub>-S<sub>8a</sub>) following syllables ending with [u] (e.g. [liea] and [luʂa]) (see Figure 34, right). Subsequently, participants were tested on whether they believed the syllables presented in isolation were the “same” or “different” from one another. Noguchi found that the Complementary

group had lower sensitivity (lower  $d'$ ) than the Control and the Non-Complementary groups. Noguchi interprets this result as showing that the Complementary group treated [ɛ] and [ɤ] as allophones of the same phoneme.

## 2.2. A ONE- OR TWO-STEP MODEL OF ALLOPHONY ACQUISITION

The tradition of studying the acquisition of phonetic categories (for example, Maye and Gerken, 2000; McGuire, 2007; Boersma et al. 2003) in isolation from the study of the acquisition of phonemes consisting of multiple phonetic categories (for example, Boersma and Hayes, 2001; Peperkamp et al., 2006) has carried on under the assumption that phonetic categories are formed before language learners form phonemes which consist of multiple phonetic categories (such as Harris, 1963; see Dillon et al., 2013 for a discussion). Peperkamp et al. (2003) explicitly propose that language learners construct phonemes in two steps: first by constructing a number of phonetic categories, and subsequently by clustering these categories into phonemes based on whether they occur in distinct contexts.

However, Dillon et al. (2013) argue that a two-step model of phoneme acquisition is not a feasible hypothesis for how language learners acquire allophonic relationships, given the large amount of category overlap exhibited across phonemes. They draw from the example of Inuktitut, which contains three vowel phonemes: /i/, /u/, and /a/. Each of these phonemes consist of two allophones: respectively [i], [u] and [a] which occur after non-uvulars, and lowered vowel qualities [e], [o], and [ɑ] which occur after uvulars. In a two-step model, the learner must discover six phonetic categories in an initial step, then determine that [i] and [e] for example occur in complementary environments and therefore are allophones of a single phoneme.

In a clustering analysis of Inuktitut vowels, Dillon and colleagues show that a machine learner performs poorly if tasked with determining the six allophones of Inuktitut. They show that a simple mixture of Gaussians model either discovers too few allophones, or discovers clusters which do not resemble actual phonetic categories well enough for learners to then determine that these categories are in complementary environments with other categories in a second step. Because of this, Dillon and colleagues

suggest a one-step model of phonological acquisition, in which allophones and rules relating allophones to one another are acquired in a single step. In their model, learners search for subsets of sets, under the condition that subsets are Gaussian-distributed and are in complementary distribution with other subsets within their set. They find that modelling with a multivariate mixture of linear models is more accurate in approximating the allophones and phonemes of Inuktitut compared to a simple mixture of Gaussians model.

The artificial language learning experiment described below will test for whether experimental evidence supports a two-stage model or a one-stage model by mapping the learning trajectory of participants exposed to one of three distribution types.

### 3. Research Question

This study has two goals. The first is to simply replicate the results of Noguchi (2016) through Mechanical Turk. In a two-day long study, Noguchi found that even after one day of training a group trained on a bimodal distribution where  $S_{1a}$ - $S_{4a}$  occurred after [i] and  $S_{5a}$ - $S_{8a}$  occurred after [u] (Bimodal-Comp group)<sup>23</sup> had significantly lower  $d'$  values than a group trained on a bimodal distribution where all tokens along the  $S_{1a}$ - $S_{8a}$  continuum occurred after [i] and after [u] (Bimodal-NonComp group). This study will also include a monomodal group for comparison.

The second and more interesting goal of this study is to determine whether there is experimental evidence for a one-stage model of phoneme acquisition or a two-stage model. In order to do so, this study will randomly place participants into one of three training conditions: **Bimodal-Comp** training, which

---

<sup>23</sup> Noguchi (2016) gives articulatory reasons that this rule, [ɛ] after [i] and [ʂ] after [u], is more natural than the opposite rule, [ɛ] after [u] and [ʂ] after [i], based on shared phonetic features between the target fricative and the contextual vowel. The articulations of palatal [ɛ] and high front vowels both require a raised tongue body, and the articulations of retroflex [ʂ] and back vowels both require a retracted tongue body. Noguchi (2016) conducts a second experiment with both a natural and an unnatural condition. Only those who were in the condition which was trained on the natural rule show decreased sensitivity compared to the Non-Complementary group; those trained on the unnatural rule showed no evidence of decreased sensitivity.

will consist of exposure to a bimodal frequency distribution where both peaks of the bimodal distribution occur in complementary environments; **Bimodal-NonComp** training, which will consist of exposure to a bimodal frequency distribution where both peaks of the bimodal distribution occur in non-complementary environments; and **Monomodal** training, which will consist of exposure to a monomodal frequency distribution. Learners will be exposed to one of three exposure times, with the greatest exposure period consisting of roughly the same number of critical tokens as that found in Noguchi’s first day of training. Noguchi found a significant difference in  $d'$  values between the Bimodal-Comp and Bimodal-NonComp groups after one day of training, so this study will use that point as roughly the last exposure time in order to determine how each group behaves before that point.

The previous chapter found a difference in sensitivity between the bimodally- and monomodally-trained participants for the [e-ɜ] critical stimuli (Experiment A3). Therefore, it is expected that the Bimodal-NonComp group will achieve a higher sensitivity than the Monomodal group at some point in time (see blue (Monomodal) and green (Bimodal-NonComp) lines in Figure 35). Additionally, based on Noguchi (2016) it is predicted that the Bimodal-NonComp group will have a significantly higher sensitivity compared to the Bimodal-Comp group after exposure to at least 256 critical syllables during training (see red (Bimodal-Comp) and green (Bimodal-NonComp) points in Figure 35).

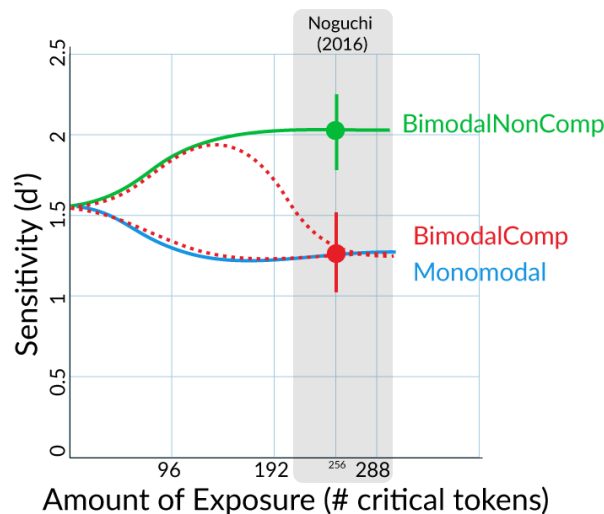


Figure 35. Predicted results as amount of exposure increases. This study will test participants trained in one of three exposure amounts, with 96 critical tokens in the shortest amount of training, and 288 critical



tokens in the greatest amount of training. The training phase of the first day of Noguchi (2016) consisted of 256 critical tokens.

The learning trajectory of interest is that of the Bimodal-Comp group. The learning trajectory of the Bimodal-Comp training is hypothesized to follow one of two trajectories, shown with the dotted lines in Figure 35. If phoneme acquisition follows a two-step model in which learners first acquire phonetic categories through distributional learning, and then learn that two phonetic categories are allophones of a single phoneme, one would expect the Bimodal-Comp group to initially pattern with the Bimodal-NonComp group, and later pattern with the Monomodal group. If phoneme acquisition follows a one-stage model of acquisition, in which learners are searching for subsets of sets from the very beginning, where each subset must be in complementary distribution with any other subsets in the same set, then the Bimodal-Comp group should always pattern with the Monomodal group. To summarize, the two research questions asked in this chapter are as follows:

- (1) Can the results of Noguchi (2016) be replicated through an online platform such as Mechanical Turk?
- (2) Does experimental evidence support a one- or a two-stage model of allophony acquisition?

This study successfully replicates Noguchi (2016) on Mechanical Turk, and finds non-significant trends which more strongly support a one-stage model of allophony acquisition over a two-stage model. Some support is also found for the model proposed in the previous chapter, that a change in bias precedes a change in sensitivity.

#### **4. Methodology**

Experiment B closely follows the methodology of Noguchi (2016). The main differences between Noguchi and the current experiment are that (1) the current study trains participants on one of three exposure amounts, (2) a Rule Test phase will be included to determine whether there is evidence that learners exposed to a complementary distribution of critical phones learned a phonological rule, and (3) a group trained on a monomodal distribution (rather than the control group used by Noguchi) will be included.

#### 4.1. STIMULI

Four types of syllables were created: critical syllables, filler syllables, context syllables, and generalization syllables. Following Noguchi (2016), onsets of critical syllables were drawn from an 8-point continuum ranging between an alveopalatal fricative [ç] to a retroflex fricative [ʂ], and each onset was followed by [ɑ]. Continuum points will be referred to as S<sub>1</sub>a-S<sub>8</sub>a, where S<sub>1</sub>a indicates the most [çɑ]-like end of the continuum, and S<sub>8</sub>a indicates the most [ʂɑ]-like end. Filler syllables consisted of four different tokens each of the syllables [tɑ] and [t<sup>h</sup>ɑ]. Context syllables were [pi pu hi hu ni nu], where each of the three onsets end with either [i] or [u]. Generalization syllables also ended in either [i] or [u], but had different onsets [ti tu fi fu li lu k<sup>h</sup>i k<sup>h</sup>u mi mu .i .u].<sup>24</sup>

All recordings, filtering, and splicing were made in Praat (version 6.0.29, Boersma, 2002), software for speech analysis, synthesis, and manipulation. Stimuli were recorded by the experimenter, a native speaker of English and heritage speaker of Mandarin. Recordings were made in a soundproof booth on an HP Spectre laptop at 44100 Hz using an ATR2500-USB Audio Technica microphone. All syllable types were recorded in two-syllable “phrases,” with context syllables and generalization syllables occurring phrase-initially, and test syllables and filler syllables occurring phrase-finally. Specifically, context and generalization syllables were followed by the dummy syllable [ʃɑ] (e.g. [li ʃɑ]), and test syllables and filler syllables were preceded by the dummy syllable [ɑ] (e.g. [ɑ çɑ]). Before manipulations were made, all recordings were high-pass filtered for frequencies that were equal to or less than 200 Hz. Dummy syllables (i.e. phrase-final [ʃɑ] and phrase-initial [ɑ]) were then spliced out. All cuts were made where the waveform crossed 0 Hz to avoid clicks and other unnatural non-speech sounds when splicing sounds together.

---

<sup>24</sup> Noguchi (2016) modelled his artificial language on Mandarin, and so used voiceless unaspirated stops (e.g. [t]) rather than voiced stops (e.g. [d]). Likewise, note that [p] is a voiceless unaspirated stop, not to be confused with [p<sup>h</sup>].

For critical syllables, tokens of the two endpoints of the target continuum [εɑ] and [ʂɑ] were recorded (again, spliced from an original recording of dummy-critical “phrases”). The fricative portions ([ε] and [ʂ]) and vowel portions ([ɑ]) were isolated. The middle 160 ms of each fricative was extracted using a parabolic windowing function. The mean intensity of each fricative was adjusted to 60 dB. To create the fricative portion of the 8-point continuum, the endpoint fricatives were overlapped in varying amounts, with the second point of the continuum consisting of 6/7ths of the [ε] token and 1/7th of the [ʂ] token, the third point of the continuum consisting of 5/7ths of the [ε] token and 2/7th of the [ʂ] token, etc. The continuum between vowels was created by using TANDEM-STRAIGHT, software which creates natural-sounding continua between two sounds. The vowel spliced from [εɑ] and the vowel spliced from [ʂɑ] were used as input into TANDEM-STRAIGHT (details of the process TANDEM-STRAIGHT uses to create continua can be found in Kawahara et al. (2008)). TANDEM-STRAIGHT returned 6 intermediate stimuli, for a total of 8 continuum points including the endpoints. All vowel continuum points were then scaled to have a mean intensity of 72 dB. Following this, each of the 8 fricative sounds were spliced onto their corresponding 8 vowel sounds, creating an 8-point continuum between [εɑ] and [ʂɑ]. Each of these syllables was scaled to have an average intensity of 74 dB. Critical syllables will be referred to as S<sub>1a</sub>-S<sub>8a</sub>, where S<sub>1a</sub> refers to the most [εɑ]-like end, and S<sub>8a</sub> refers to the most [ʂɑ]-like end. Spectrograms of critical syllables S<sub>1a</sub>, S<sub>3a</sub>, S<sub>6a</sub>, and S<sub>8a</sub> are shown in Figure 26. Note that these critical syllables are identical to those used in Experiments A3-1 and A3-2 from the previous chapter.

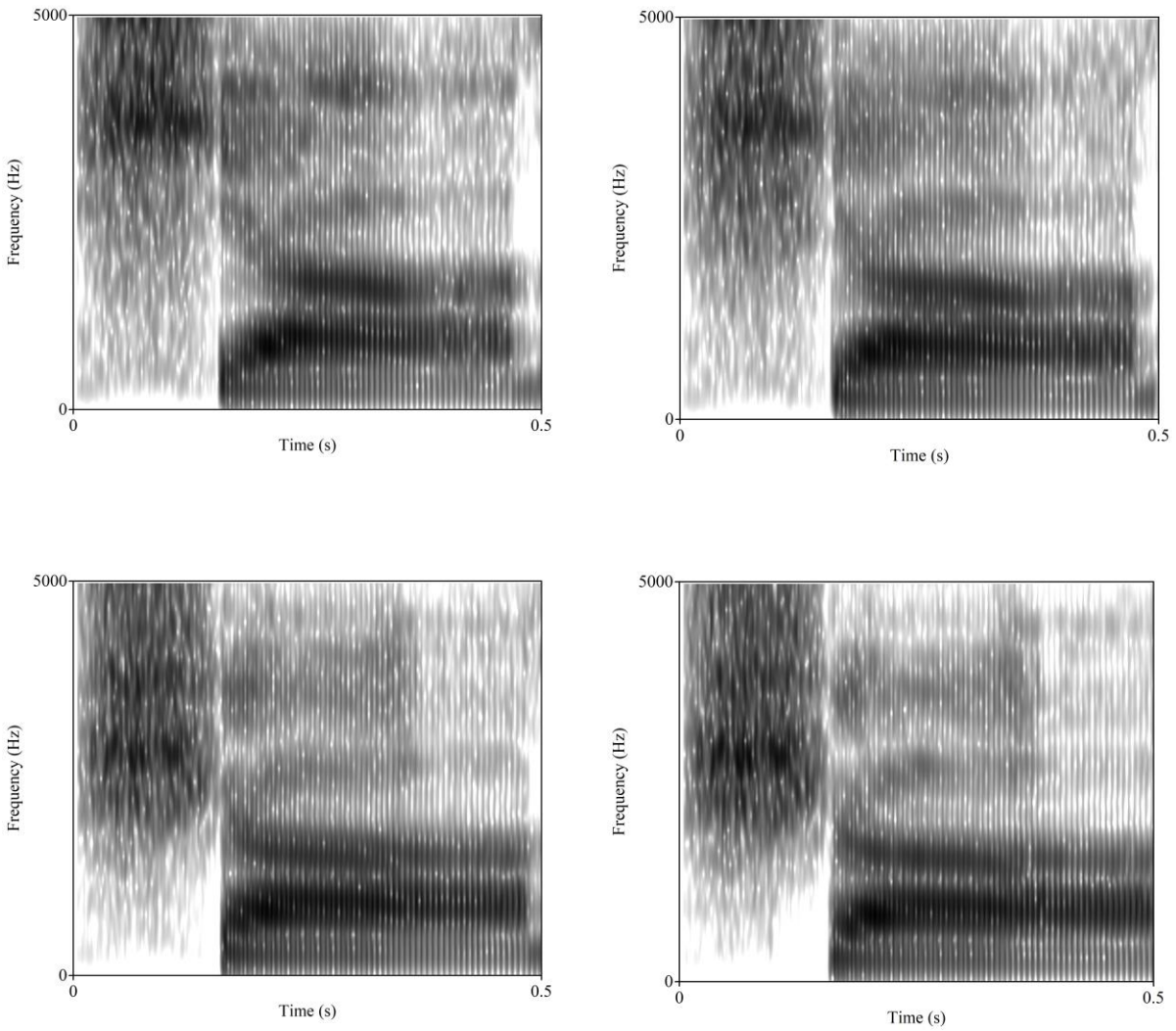


Figure 36. First 500 ms of critical syllables  $S_{1a}$  (top left),  $S_{3a}$  (top right),  $S_{6a}$  (bottom left), and  $S_{8a}$  (bottom right).

Each of the 6 context syllables were concatenated before each of the 8 critical syllables and each of the 8 filler syllables. This made up the stimuli to be used during training. Each of the 12 generalization syllables were concatenated before  $S_{1a}$  and  $S_{8a}$ , and before each of the 8 filler syllables. These stimuli were used in the Rule Test, which is described below.

## 4.2. PROCEDURE

Participants were randomly placed into one of three Distributions: Bimodal-Comp, Bimodal-NonComp, or Monomodal. Participants were also randomly placed into one of three ExposureTimes (One, Two, or Three), and one of two TestOrders (RuleFirst or PhoneFirst). This experiment consisted of five parts: a Practice Phone Test in English, followed by a Train phase, followed by two tests, a Phone Test and a Rule Test, followed by a Questionnaire. Participants were directed to either the Phone Test first or the Rule Test first depending on which TestOrder condition they were in.

<b>Procedure</b>	<b>Abbreviated directions</b>	<b>Sample trial stimuli</b>
1. Practice phone test	Are these two words the same or different?	<i>sheep</i> vs. <i>ship</i> <i>sheep</i> <sub>1</sub> vs. <i>sheep</i> <sub>2</sub>
2. Training	Listen carefully	<i>ni S<sub>1</sub>a</i> <i>ni ta</i>
3. Phone Test	Are these two words the same or different?	<i>S<sub>1</sub>a</i> vs. <i>S<sub>8</sub>a</i> <i>S<sub>1</sub>a</i> vs. <i>S<sub>1</sub>a</i>
3. Rule Test	Which of these two phrases are allowed in this language?	<i>ni S<sub>1</sub>a</i> vs. <i>ni S<sub>1</sub>a</i>
5. Questionnaire	Please provide background information (responses will not affect payment)	

Table 17. Summary of procedure in Experiment B.

At the beginning of the experiment, participants were given the following instructions:

*[Page 1]*

*Just a few things to keep in mind before you begin!  
Please wear headphones for the duration of this experiment...  
Please do not write any words down while taking this experiment...  
...And please do not click on your browser's back or refresh button.*

*[Page 2]*

*A few more things to keep in mind before you begin...*

*As this is a scientific experiment, it is important that you devote your **full attention** to this HIT<sup>25</sup>! Please do not do other tasks or shrink the browser window, and please do not remove your headphones.*

---

<sup>25</sup> A “HIT” (Human Intelligence Task) is a term used in Mechanical Turk, referring to a task that a participant (“Worker”) can complete.

*If you are unable to devote your full attention to this HIT, which is expected to take about 20-40 minutes, **please do not do this HIT.***

Participants were then directed to the English Practice Phone Test.<sup>26</sup> They were given the following instructions:

*In this English practice test, you will see two buttons on the screen: one labelled “**SAME**”, and one labelled “**DIFFERENT**”.*

*You will hear someone saying two things in English. They are either two repetitions of the same word or two different words. If you think they are repetitions of the same word, press the “**SAME**” button. If you think they are different words, press the “**DIFFERENT**” button.*

During the English Practice Phone Test, participants were given a pair of English words, such as *sheep* and *ship*. Participants were asked to determine whether the second words in these phrases were the same word, or two different words. Participants were given 4 same pairs (e.g. *ship*<sub>1</sub> vs. *ship*<sub>2</sub>), and 4 different pairs (e.g. *ship*<sub>1</sub> vs. *sheep*<sub>1</sub>). No feedback was given.

After completing the Practice Phone Test, participants were directed to a Train phase. They were given the following instructions:

*[Page 1]*

*Great job! Later, we will do the same thing with a **foreign language**. First though, you will listen to short phrases in this foreign language. Each recording consists of two words. Words in this language are very short and consist of only one syllable. Therefore a phrase could be something like “wa ko”.*

*This will take 10-20 minutes. Your main job is to listen carefully.*

*[Page 2]*

*Please remember: **DO NOT WRITE ANY WORDS DOWN**. You may draw or sketch while you are listening to pass the time, but please do not write any words down. This section will feel very long since you are listening passively for the most part, but in reality takes no more than about 20 minutes.*

---

<sup>26</sup> Note that there was no Sound Check phase in this experiment, in order to keep the experiment from being too long and avoid fatigue for participants in ExposureTime Three.

Exposure items presented during the training phase consisted of context syllables followed by critical syllables (e.g. *ni S<sub>1a</sub>*), or context syllables followed by filler syllables (e.g. *ni ta*).

Participants in the Bimodal-Comp and Bimodal-NonComp groups were exposed to critical phones whose frequencies fell in a bimodal distribution, and participants in the Monomodal group were exposed to critical phones whose frequencies fell in a monomodal distribution. Bimodal-NonComp and Monomodal groups were exposed to all continuum points S<sub>1a</sub>-S<sub>8a</sub> after [i] and [u], whereas the Bimodal-Comp group only heard S<sub>1a</sub>-S<sub>4a</sub> (the [ɛ]-like end of the continuum) after [i], and S<sub>5a</sub>-S<sub>8a</sub> (the [ʂ]-like end of the continuum) after [u] (Noguchi, 2016). This is articulatorily more natural than the opposite pattern, S<sub>1a</sub>-S<sub>4a</sub> after [u], and S<sub>5a</sub>-S<sub>8a</sub> after [i], as palatal consonants and high front vowels both require a raised tongue body, and retroflex consonants and back vowels require a retracted tongue body. Figure 37 illustrates the frequency distributions participants were exposed to.

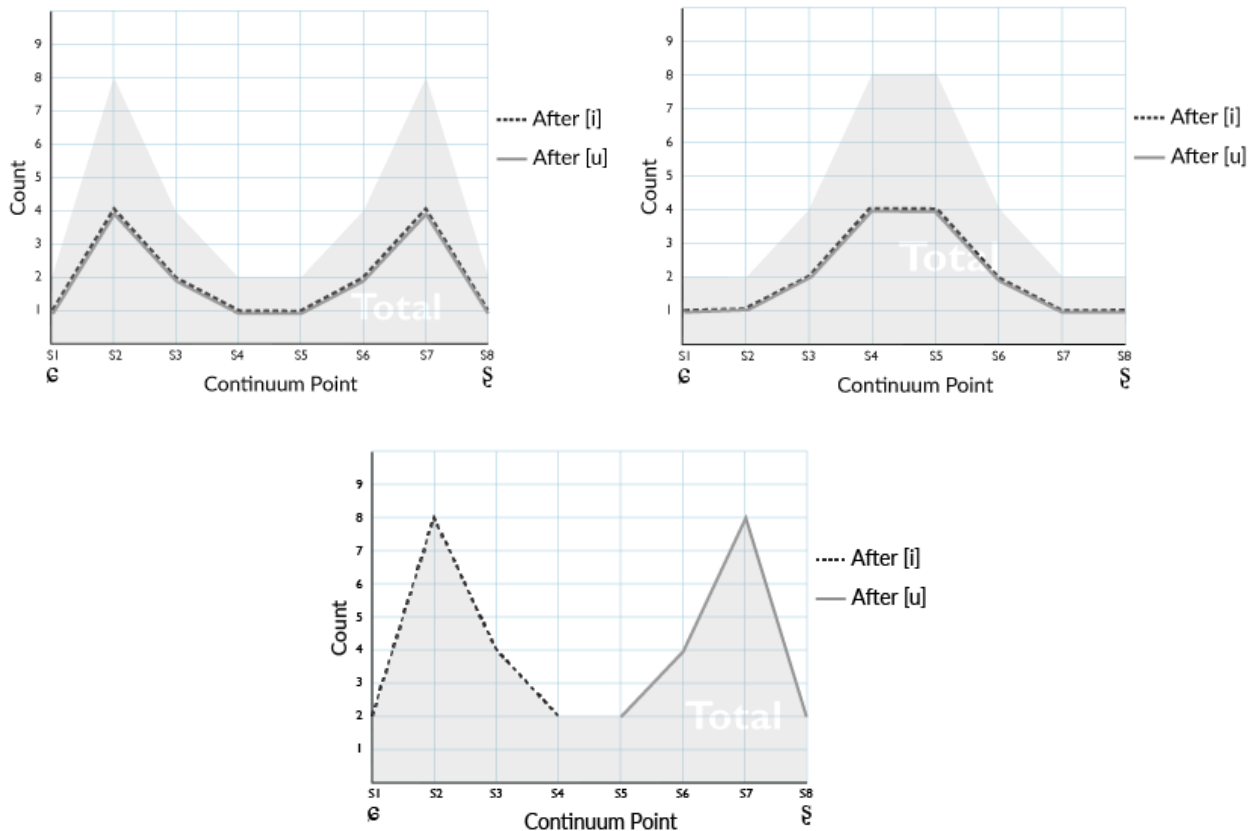


Figure 37. Illustration of familiarization frequency for Bimodal-NonComp group (top-left), Monomodal group (top-right), and Bimodal-Comp group (bottom) during training.

Each block of training consisted of one of each of the 6 context syllables [pi pu hi hu ni nu] followed by 16 critical syllables S<sub>1a</sub>-S<sub>8a</sub> (following the distributions shown in Figure 3), resulting in 96 critical phrases per block. Each block of training also consisted of one of each of the 6 context syllables followed by one of the 4 tokens of 2 filler syllables, resulting in 48 fillers per block. Participants in the ExposureTime One group were exposed to one block of training stimuli (96 critical stimuli), which took about 5 minutes; participants in the ExposureTime Two group were exposed to two blocks of training stimuli (192 critical stimuli), which took about 10 minutes; and participants in the ExposureTime Three group were exposed to three blocks of training stimuli (288 critical stimuli), which took about 15 minutes.

After training, participants were directed to one of the two test phases, the order of which depended on which TestOrder participants were in. Before the Phone Test, participants were given the following instructions:

*Great job! This next part will be similar to the practice test you did earlier in English, but this time it will ask you about the foreign language that you heard.*

*Again, you will see two buttons on the screen: a “SAME” button, and a “DIFFERENT” button. You will hear a person saying two things in the foreign language you heard. If you think they are repetitions of the same word in this language, press the “SAME” button. If you think they are different words, press the “DIFFERENT” button.*

In the Phone Test, participants were presented with pairs of syllables. The Phone Test consisted of 12 critical Different Pairs (S<sub>1a</sub> vs. S<sub>8a</sub>), 12 critical Same Pairs (S<sub>1a</sub> vs. S<sub>1a</sub> or S<sub>8a</sub> vs. S<sub>8a</sub>), 12 filler Different Pairs ([ta]<sub>1</sub> vs. [tʰa]<sub>1</sub>), and 12 filler Same Pairs ([ta]<sub>1</sub> vs. [ta]<sub>2</sub>).

The purpose of the Rule Test was to determine whether participants had learned the phonological rule that [ɛ] occurs after [i] and [ɤ] occurs after [u], and, if so, whether they had generalized this rule to apply to newly-heard syllables. Before the Rule Test, participants were given the following instructions:

*Great job! In this next section, you will hear a short phrase in this foreign language. You will be asked whether the phrase is allowed in the language you heard.*

*Please note that you have not heard all of the phrases that are allowed in this language. If you are not sure if the phrase is allowed or not, just select your best guess.*



In the Rule Test, participants heard one two-syllable phrase and were given two button options to click on. One button read *This phrase IS allowed*, and the other read *This phrase is NOT allowed*. Trials in the rule test consisted of either critical trials or filler trials. In critical trials, participants heard either S<sub>1</sub>a or S<sub>8</sub>a preceded by either a context syllable [pi pu hi hu ni nu] which they had heard during training (“old trial”), or a new generalization syllable [ti tu fi fu li lu k<sup>h</sup>i k<sup>h</sup>u mi mu .i .u] which they had not heard during training (“new trial”). Critical trials were either Legal trials, which conformed to the rule that S<sub>1</sub>a follows [i] and S<sub>8</sub>a follows [u], or Illegal trials, which violated this rule. Filler trials consisted of a generalization syllable followed by either [ta] or [t<sup>h</sup>a]. Participants heard 24 old trials (12 Legal, 12 Illegal)<sup>27</sup>, 24 new trials (12 Legal, 12 Illegal), and 24 filler trials.

Participants in the PhoneFirst condition were presented with the Phone Test first, followed by the Rule Test. Participants in the RuleFirst condition took the Rule Test first, followed by the Phone Test.

After the experiment, participants were directed to a short questionnaire which asked about participants’ demographic information (age, place of residence, etc.), language background (native language, languages studied, history of speech or hearing disorder, etc.), and attention levels during participation. Participants were also asked whether they had noticed any patterns and whether they used any strategies during the experiment.

#### 4.3. PARTICIPANTS

Participants were asked to participate only if they (1) had no known history of speech or hearing impairments, (2) were a native speaker of English, (3) had regular access to a computer with an internet connection, and (4) were using a computer able to play audio. Because this experiment was conducted online rather than face-to-face, only participants using a computer in the United States were allowed to

---

<sup>27</sup> Note that “Old” and “New” refer to the familiarity of the syllable, and not of the phrase. Therefore, an “Old Illegal” trial is a trial which consists of an Old context syllable (one used during Training) followed by a critical syllable, where the entire “phrase” violates the rule that S<sub>1</sub>a follows [i] and S<sub>8</sub>a follows [u].

participate to increase the chance that the participant would be a native English speaker using an MTurk qualification (attributes that participants on MTurk can obtain). In addition, since the onsets of the critical syllables ranged between [ɛ] and [ʃ], following Noguchi (2016), participant responses were not included in analysis if they reported having some background in a language with more than one voiceless post-alveolar fricative as phonemes. Participants were asked to not participate if they had some language background in Mandarin Chinese, Japanese, Russian, or German. Qualifications used to screen participants are as follows:

- Must be using a computer in the United States
- Must have an approval rating of equal or greater to 90% on all tasks completed on MTurk
- Must have had at least 50 completed MTurk tasks approved

431 participants were recruited through Mechanical Turk. Participants were excluded if they: 1) scored fewer than 5/8 correct on the practice English test (15 excluded for this); 2) reported having a speech or hearing disorder in the questionnaire (4 excluded for this); 3) reported not being a native English speaker (1 excluded for this); or 4) reported having some sort of back-ground with a language with more than one voiceless post-alveolar fricative (22 excluded for this). The number of participants analyzed per condition are displayed in Table 18 (note that some participants were excluded for more than one reason).

		ExposureTime One	ExposureTime Two	ExposureTime Three
Bimodal-Comp	PhoneFirst	23	23	25
	RuleFirst	18	17	18
	<b>Total</b>	<b>41</b>	<b>40</b>	<b>43</b>
Bimodal-NonComp	PhoneFirst	31	17	19
	RuleFirst	20	16	21
	<b>Total</b>	<b>51</b>	<b>33</b>	<b>40</b>
Monomodal	PhoneFirst	28	23	26
	RuleFirst	20	21	25
	<b>Total</b>	<b>48</b>	<b>44</b>	<b>51</b>

Table 18. Number of participants included in analysis per condition.

The number of participants that were included in the analysis ranged between 33 in the Bimodal-NonComp, ExposureTime Two condition, to 51 in the Bimodal-NonComp, ExposureTime One and Monomodal, ExposureTime Three condition.

## 5. Results

This section first describes the results of the Phone Test (Section 5.1), followed by results of the Rule Test (Section 5.2).

### 5.1. PHONE TEST

The Phone Test will be analyzed the same way the Test results from Chapter 3 were analyzed, but will include a third independent factor, ExposureTime. As with Chapter 3, this section will be adhering to the following conventions to interpret findings:

- (1) A significant interaction between condition and stimulus type will be interpreted as a significant difference in **sensitivity** between conditions
- (2) A significant main effect of condition will be interpreted as a significant difference in **response bias** between conditions...
- (3) ... *unless* a significant interaction between condition and stimulus type was also found, in which case a main effect will not be interpreted.

Again, results of this *same-different* test will be treated as *no-yes* or *noSignal-signal* experiments (see Chapter 2).

The regression used in analysis modelled one dependent variable, Response, with three fixed effects: (1) **Distribution**, consisting of three levels {*bimodal-comp*, *bimodal-nonComp*, *monomodal*}, (2) **PairType**, consisting of two levels {*same*, *diff*}, and (3) **ExposureTime**, consisting of three levels {*one*, *two*, *three*}. The dependent variable **Response** consists of two levels, *s* and *d*, where *s* corresponds to a participant response of “same” during the Test phase, and where *d* corresponds to a participant response of “different” during the Test phase. The regression formula initially used is shown in (1).

- (1) 
$$\text{Response} \sim \text{Condition} * \text{PairType} * \text{ExposureTime} + (1 + \text{PairType} | \text{Subject}) + (1 + \text{ExposureTime} + \text{Condition} | \text{Item})$$

Due to a failure to converge, the random effects structure was simplified (see Barr et al., 2013). The final formula the regression was fitted to included random slopes by Subject and by Item, as shown in (2).

- (2) 
$$\text{Response} \sim \text{Distribution} * \text{PairType} * \text{ExposureTime} + (1 | \text{Subject}) + (1 | \text{Item})$$

Results of the model for critical and filler stimuli are shown in Table 19.

<b>Predictor</b>	<b>Coefficient</b>	<b>SE</b>	<b>Wald Z</b>	<b>p</b>
<b>CRITICAL</b>				
(Intercept)	-0.095	0.328	-0.288	0.773
Distribution= <i>bimodalNonComp</i>	0.639	0.436	1.467	0.142
Distribution= <i>monomodal</i>	-0.186	0.442	-0.42	0.675
PairType= <i>same</i>	-4.230	0.309	-13.686	<0.001 ***
ExposureTime= <i>three</i>	0.006	0.452	0.013	0.990
ExposureTime = <i>two</i>	0.445	0.458	0.972	0.331
Interaction= <i>bimodalNonComp &amp; same</i>	-0.003	0.383	-0.008	0.993
Interaction= <i>monomodal &amp; same</i>	-0.497	0.443	-1.121	0.262
Interaction= <i>bimodalNonComp &amp; three</i>	-0.270	0.629	-0.428	0.668
Interaction= <i>monomodalNonComp &amp; three</i>	0.330	0.617	0.534	0.593
Interaction= <i>bimodalNonComp &amp; two</i>	-0.672	0.652	-1.031	0.303
Interaction= <i>monomodalNonComp &amp; two</i>	-0.431	0.634	-0.68	0.496
Interaction= <i>same &amp; three</i>	0.672	0.387	1.736	0.083
Interaction= <i>same &amp; two</i>	0.498	0.387	1.287	0.198
Interaction= <i>bimodalNonComp &amp; same &amp; three</i>	-0.903	0.547	-1.65	0.099
Interaction= <i>monomodalNonComp &amp; same &amp; three</i>	0.002	0.560	0.004	0.997
Interaction= <i>bimodalNonComp &amp; same &amp; two</i>	-0.143	0.533	-0.267	0.789
Interaction= <i>monomodalNonComp &amp; same &amp; two</i>	-0.334	0.599	-0.558	0.577
<b>FILLER</b>				
(Intercept)	2.756	0.217	12.685	<0.001 ***
Distribution= <i>BimodalNonComp</i>	-0.121	0.268	-0.453	0.651
Distribution= <i>Monomodal</i>	-0.201	0.269	-0.746	0.456
PairType= <i>same</i>	-5.271	0.270	-19.51	<0.001 ***
Timepoint= <i>Three</i>	0.199	0.293	0.680	0.496
Timepoint= <i>Two</i>	-0.254	0.278	-0.913	0.361
Interaction= <i>bimodalNonComp &amp; same</i>	0.125	0.325	0.385	0.700
Interaction= <i>monomodal &amp; same</i>	0.096	0.330	0.290	0.772
Interaction= <i>bimodalNonComp &amp; three</i>	0.054	0.403	0.135	0.893
Interaction= <i>monomodalNonComp &amp; three</i>	0.098	0.390	0.252	0.801
Interaction= <i>bimodalNonComp &amp; two</i>	0.226	0.393	0.574	0.566
Interaction= <i>monomodalNonComp &amp; two</i>	0.433	0.382	1.132	0.258
Interaction= <i>same &amp; three</i>	-0.277	0.352	-0.785	0.433
Interaction= <i>same &amp; two</i>	-0.087	0.351	-0.247	0.805
Interaction= <i>bimodalNonComp &amp; same &amp; three</i>	-0.160	0.489	-0.328	0.743
Interaction= <i>monomodalNonComp &amp; same &amp; three</i>	-0.144	0.477	-0.302	0.763
Interaction= <i>bimodalNonComp &amp; same &amp; two</i>	0.585	0.479	1.220	0.222
Interaction= <i>monomodalNonComp &amp; same &amp; two</i>	-0.132	0.480	-0.274	0.784

Table 19. Results of GLMM for the phone test.

Follow-up contrasts in the context of the overall model were completed to test the following hypotheses:

- The interaction between Distribution and PairType is significant for the Bimodal-Comp group compared to the Bimodal-NonComp group
- The interaction between Distribution and PairType is significant for the Bimodal-Comp group compared to the Monomodal group
- The interaction between Distribution and PairType is significant for the Bimodal-NonComp group compared to the Monomodal group

These three hypotheses were tested at each of the three ExposureTimes. Results are summarized in Table 20.

ExposureTime	Distribution comparison	Coefficient	SE	Wald Z	<i>p</i>
<b>CRITICAL</b>					
One	Bimodal-NonComp vs. Bimodal-Comp	-0.003	0.383	-0.008	0.993
	Monomodal vs. Bimodal-Comp	-0.497	0.443	-1.121	0.262
	Monomodal vs. Bimodal-NonComp	-0.494	0.405	-1.220	0.223
Two	Bimodal-NonComp vs. Bimodal-Comp	-0.146	0.372	-0.392	0.695
	Monomodal vs. Bimodal-Comp	-0.831	0.403	-2.060	0.039 *
	Monomodal vs. Bimodal-NonComp	-0.685	0.424	-1.614	0.106
Three	Bimodal-NonComp vs. Bimodal-Comp	-0.906	0.391	-2.316	0.021 *
	Monomodal vs. Bimodal-Comp	-0.494	0.343	-1.442	0.149
	Monomodal vs. Bimodal-NonComp	0.411	0.387	1.063	0.288
<b>FILLER</b>					
One	Bimodal-NonComp vs. Bimodal-Comp	0.125	0.325	0.385	0.700
	Monomodal vs. Bimodal-Comp	0.096	0.330	0.290	0.772
	Monomodal vs. Bimodal-NonComp	-0.030	0.308	-0.096	0.923
Two	Bimodal-NonComp vs. Bimodal-Comp	0.710	0.352	2.017	0.044 *
	Monomodal vs. Bimodal-Comp	-0.036	0.349	-0.103	0.918
	Monomodal vs. Bimodal-NonComp	-0.746	0.344	02.168	0.030 *
Three	Bimodal-NonComp vs. Bimodal-Comp	-0.035	0.365	-0.096	0.924
	Monomodal vs. Bimodal-Comp	-0.048	0.344	-0.140	0.889
	Monomodal vs. Bimodal-NonComp	-0.013	0.351	-0.038	0.970

Table 20. Summary of follow-up contrasts testing specific hypotheses for the Phone Test.

At ExposureTime Two, the interaction between Condition and PairType when comparing the Bimodal-Comp group with the Monomodal group is significant for critical trials ( $p = 0.039$ ), with those exposed to a Bimodal-Comp distribution having 0.831 lesser log-odds of responding “different” than those exposed to a Monomodal distribution at that ExposureTime. At ExposureTime Three, the interaction between Condition and PairType when comparing the Bimodal-Comp group with the Bimodal-NonComp group is significant for critical trials ( $p = 0.021$ ), with participants exposed to a Bimodal-Comp distribution having -0.906 lesser log-odds of responding “different” than those exposed to a Bimodal-NonComp distribution. There are no significant interactions between Condition and PairType at ExposureTime One. Sensitivity at each ExposureTime is shown in Figure 38.

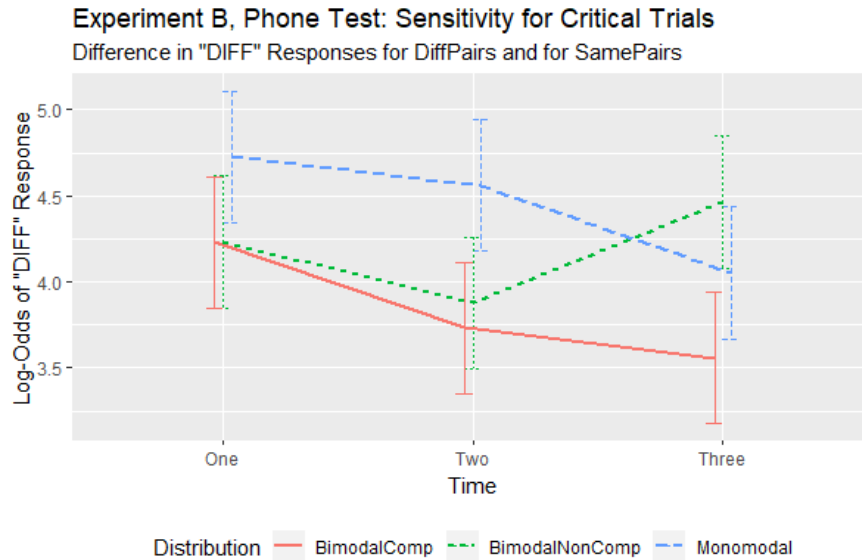


Figure 38. The difference (in log-odds) of participants responding that critical Same Pairs are “different” and responding that critical Different Pairs are “different.”

Findings are interpreted as follows: with the least amount of exposure tested in this experiment (ExposureTime 1), no groups differ significantly from one another in terms of sensitivity. At the second-most amount of exposure tested in this experiment (ExposureTime 2), the Bimodal-Comp group and the Monomodal group do not pattern together in terms of sensitivity, with the Monomodal group having 0.831 more sensitivity (measured in log-odds) than to the Bimodal-Comp group. At ExposureTime 3, the Bimodal-Comp and Bimodal-NonComp groups do not pattern together, with the Bimodal-NonComp group having 0.906 more sensitivity (measured in log-odds) than the Bimodal-Comp group, successfully replicating findings from Noguchi (2016), who finds that a Bimodal-Comp group has lowered sensitivity compared to a Bimodal-NonComp group.

For filler trials, there was a significant interaction between Condition and PairType between the Bimodal-Comp and Bimodal-NonComp groups at ExposureTime Two ( $p = 0.044$ ), as well as a significant interaction between the Monomodal and Bimodal-NonComp groups at ExposureTime Two ( $p = 0.030$ ). It is unclear what resulted in the significantly lower sensitivity for the Bimodal-NonComp group in filler stimuli at the second ExposureTime.

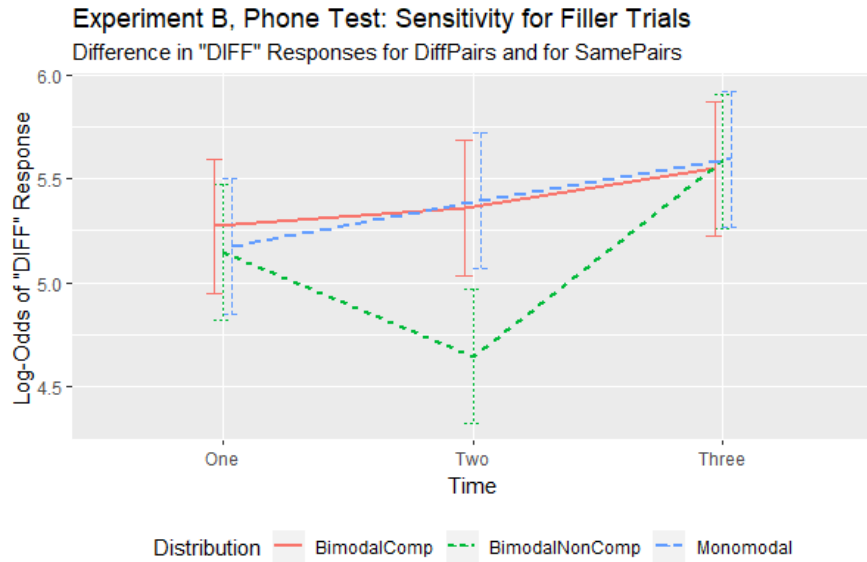


Figure 39. The difference (in log-odds) of participants responding that filler SamePairs are “different” and responding that filler DiffPairs are “different.”

Because the previous chapter argued for a model in which a change in bias precedes a change in sensitivity, main effects of Distribution are also tested for. Tests were done in the context of the overall model (that is, those results shown in Table 19). Results are shown in Table 21.

<b>ExposureTime</b>	<b>Distribution Comparison</b>	<b>Estimate</b>	<b>SE</b>	<b>Wald Z</b>	<b>p</b>
<b>CRITICAL</b>					
One	Bimodal-NonComp vs. Bimodal-Comp	-0.638	0.441	-1.444	0.149
	Monomodal vs. Bimodal-Comp	0.434	0.464	0.936	0.349
	Monomodal vs. Bimodal-NonComp	1.072	0.434	2.471	0.014 *
Two	Bimodal-NonComp vs. Bimodal-Comp	0.106	0.485	0.218	0.827
	Monomodal vs. Bimodal-Comp	1.032	0.465	2.223	0.026 *
	Monomodal vs. Bimodal-NonComp	0.927	0.490	1.890	0.059
Three	Bimodal-NonComp vs. Bimodal-Comp	0.083	0.462	0.181	0.857
	Monomodal vs. Bimodal-Comp	0.103	0.434	0.237	0.812
	Monomodal vs. Bimodal-NonComp	0.020	0.447	0.044	0.965
<b>FILLER</b>					
One	Bimodal-NonComp vs. Bimodal-Comp	0.059	0.202	0.290	0.772
	Monomodal vs. Bimodal-Comp	0.153	0.205	0.744	0.457
	Monomodal vs. Bimodal-NonComp	0.094	0.192	0.489	0.625
Two	Bimodal-NonComp vs. Bimodal-Comp	-0.460	0.223	-2.06	0.040 *
	Monomodal vs. Bimodal-Comp	-0.214	0.215	-0.994	0.320
	Monomodal vs. Bimodal-NonComp	0.245	0.218	1.125	0.261
Three	Bimodal-NonComp vs. Bimodal-Comp	0.084	0.222	0.379	0.704
	Monomodal vs. Bimodal-Comp	0.126	0.210	0.602	0.547
	Monomodal vs. Bimodal-NonComp	0.042	0.214	0.195	0.845

Table 21. Results of hypothesis testing within the context of the overall model for main effect of Distribution (regardless of PairType) for the Phone Test. Rows in grey are not interpreted since these comparisons yielded significant interactions between Distribution and PairType (see Table 20).

A significant main effect of Distribution at ExposureTime One when the Bimodal-NonComp and Monomodal groups are compared is found, with the Bimodal-NonComp group having 1.072 greater log-odds of responding “different” than the Monomodal group, regardless of PairType. This supports what was found in the previous chapter: exposure to a bimodal (non-complementary) distribution led to greater log-odds of a “different” response compared to exposure to a monomodal distribution of phones. Figure 40 summarizes the results of the follow-up contrasts testing for main effects for critical and filler stimuli.



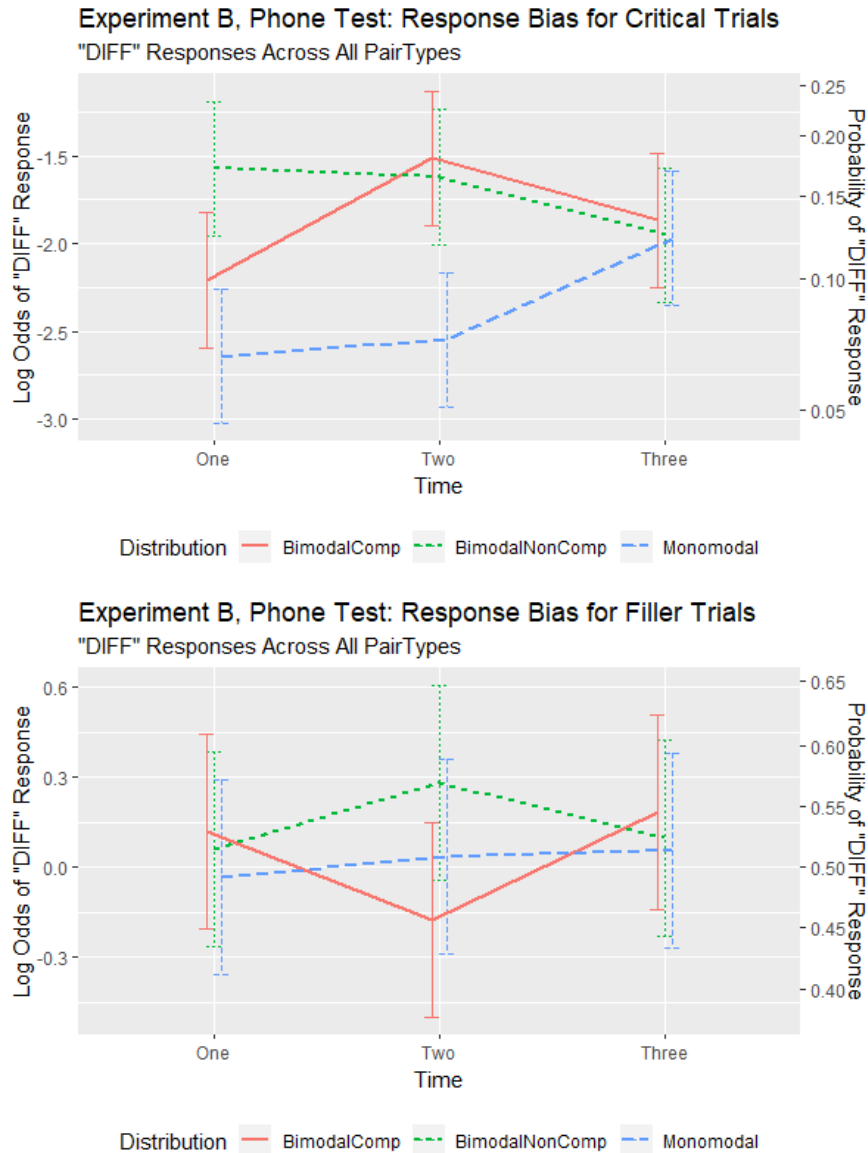


Figure 40. Main effect of Distribution for critical trials (top) and filler trials (bottom).

Two other significant main effects of Distribution are also found, but are not interpreted here. This is due to the fact that these same comparisons showed a significant interaction between PairType and Distribution. See Chapter 2, Section 3.4 for a detailed explanation as to why these cases are not interpreted.

## 5.2. RULE TEST

This section reports on results of the Rule Test, which tests whether participants 1) learn the rule that [ɛ] occurs after [i] and [ʒ] occurs after [u], and 2) generalize this rule to syllables they were not exposed to during training.

The regression used in the analysis of results of the rule test modelled one dependent variable, *RuleResponse*, with three fixed effects: (1) **Distribution**, consisting of three levels {*bimodal-comp*, *bimodal-nonComp*, *monomodal*}, (2) **TrialType**, consisting of two levels {*legal*, *illegal*}, and (3) **ExposureTime**, consisting of three levels {*one*, *two*, *three*}. The dependent variable **RuleResponse** consists of two levels, *l* and *i*, where *l* corresponds to a participant pressing the button labelled *This phrase IS allowed* (or “legal”) during the rule test phase, and where *i* corresponds to a participant pressing the button labelled *This phrase is NOT allowed* (or is “illegal”). The regression formula initially used is shown in (3).

$$(3) \quad \text{RuleResponse} \sim \text{Condition} * \text{TrialType} * \text{ExposureTime} + (1 + \text{TrialType} | \text{Subject}) + (1 + \text{ExposureTime} + \text{Condition} | \text{Item})$$

Due to a failure to converge, the random effects structure was simplified (see Barr et al., 2013). The final formula the regression was fitted to included random slopes by *Subject* and by *Item*, as shown in (4).

$$(4) \quad \text{RuleResponse} \sim \text{Distribution} * \text{TrialType} * \text{ExposureTime} + (1 | \text{Subject}) + (1 | \text{Item})$$

The regression was fitted once to *old* trials, and once to *new* trials. Results are summarized in Table 22.

Predictor	Coefficient	SE	Wald Z	p
<b>OLD</b>				
(Intercept)	1.603	0.256	6.268	<0.001 ***
Distribution= <i>bimodalNonComp</i>	0.203	0.330	0.616	0.538
Distribution= <i>monomodal</i>	0.374	0.336	1.113	0.266
TrialType= <i>legal</i>	0.209	0.199	1.047	0.295
ExposureTime= <i>three</i>	0.319	0.344	0.927	0.354
ExposureTime = <i>two</i>	-0.168	0.344	-0.487	0.627
Interaction= <i>bimodalNonComp &amp; legal</i>	-0.574	0.229	-2.506	0.012 *
Interaction= <i>monomodal &amp; legal</i>	-0.239	0.236	-1.010	0.313
Interaction= <i>bimodalNonComp &amp; three</i>	0.134	0.484	0.277	0.782
Interaction= <i>monomodalNonComp &amp; three</i>	-0.096	0.473	-0.204	0.839
Interaction= <i>bimodalNonComp &amp; two</i>	-0.086	0.491	-0.175	0.861
Interaction= <i>monomodalNonComp &amp; two</i>	0.063	0.478	0.132	0.895
Interaction= <i>legal &amp; three</i>	-0.133	0.242	-0.548	0.584
Interaction= <i>legal &amp; two</i>	0.233	0.239	0.975	0.330
Interaction= <i>bimodalNonComp &amp; legal &amp; three</i>	0.536	0.346	1.550	0.121
Interaction= <i>monomodalNonComp &amp; legal &amp; three</i>	0.299	0.339	0.883	0.377
Interaction= <i>bimodalNonComp &amp; legal &amp; two</i>	0.200	0.337	0.592	0.554
Interaction= <i>monomodalNonComp &amp; legal &amp; two</i>	-0.307	0.334	-0.916	0.360
<b>NEW</b>				
(Intercept)	0.602	0.265	2.275	0.023 *
Distribution= <i>bimodalNonComp</i>	-0.256	0.278	-0.922	0.357
Distribution= <i>monomodal</i>	-0.035	0.283	-0.125	0.901
TrialType= <i>legal</i>	-0.151	0.274	-0.550	0.582
ExposureTime= <i>three</i>	-0.414	0.291	-1.426	0.154
ExposureTime = <i>two</i>	-0.157	0.295	-0.530	0.596
Interaction= <i>bimodalNonComp &amp; legal</i>	0.293	0.194	1.508	0.132
Interaction= <i>monomodal &amp; legal</i>	0.236	0.200	1.182	0.237
Interaction= <i>bimodalNonComp &amp; three</i>	0.275	0.403	0.681	0.496
Interaction= <i>monomodalNonComp &amp; three</i>	0.028	0.395	0.070	0.944
Interaction= <i>bimodalNonComp &amp; two</i>	0.013	0.419	0.030	0.976
Interaction= <i>monomodalNonComp &amp; two</i>	-0.261	0.405	-0.645	0.519
Interaction= <i>legal &amp; three</i>	0.180	0.205	0.879	0.380
Interaction= <i>legal &amp; two</i>	0.001	0.208	0.006	0.996
Interaction= <i>bimodalNonComp &amp; legal &amp; three</i>	-0.260	0.284	-0.915	0.360
Interaction= <i>monomodalNonComp &amp; legal &amp; three</i>	-0.243	0.278	-0.873	0.383
Interaction= <i>bimodalNonComp &amp; legal &amp; two</i>	-0.337	0.294	-1.144	0.253
Interaction= <i>monomodalNonComp &amp; legal &amp; two</i>	-0.069	0.286	-0.243	0.808

Table 22. Results of GLMM for the rule test.

Follow-up contrasts in the context of the overall model were completed to test the following hypotheses:

- The interaction between Distribution and TrialType is significant for the Bimodal-Comp group compared to the Bimodal-NonComp group
- The interaction between Distribution and TrialType is significant for the Bimodal-Comp group compared to the Monomodal group
- The interaction between Distribution and TrialType is significant for the Bimodal-NonComp group compared to the Monomodal group

These three hypotheses were tested at each of the three ExposureTimes. Results of the follow-up contrasts are shown in Table 23. The difference in an “allowed” response (measured in log-odds) between *legal* and *illegal* trials is shown in Figure 41 (old trials) and Figure 42 (new trials). The higher the number, the higher the sensitivity is to the rule that [ɛɑ] occurs after [i] and [ʂɑ] occurs after [u].

ExposureTime	Distribution Comparison	Coefficient	SE	Wald Z	p
OLD					
One	Bimodal-NonComp vs. Bimodal-Comp	0.574	0.229	2.506	0.012 *
	Monomodal vs. Bimodal-Comp	0.239	0.236	1.010	0.312
	Monomodal vs. Bimodal-NonComp	-0.336	0.225	-1.495	0.135
Two	Bimodal-NonComp vs. Bimodal-Comp	0.375	0.248	1.512	0.130
	Monomodal vs. Bimodal-Comp	0.545	0.237	2.304	0.021 *
	Monomodal vs. Bimodal-NonComp	0.171	0.247	0.690	0.490
Three	Bimodal-NonComp vs. Bimodal-Comp	0.0385	0.259	0.148	0.882
	Monomodal vs. Bimodal-Comp	-0.061	0.243	-0.249	0.803
	Monomodal vs. Bimodal-NonComp	-0.099	0.258	-0.384	0.701
NEW					
One	Bimodal-NonComp vs. Bimodal-Comp	-0.293	0.194	-1.508	0.131
	Monomodal vs. Bimodal-Comp	-0.236	0.200	-1.182	0.237
	Monomodal vs. Bimodal-NonComp	0.057	0.189	0.302	0.763
Two	Bimodal-NonComp vs. Bimodal-Comp	0.044	0.221	0.198	0.843
	Monomodal vs. Bimodal-Comp	-0.166	0.205	-0.813	0.416
	Monomodal vs. Bimodal-NonComp	-0.210	0.215	-0.976	0.329
Three	Bimodal-NonComp vs. Bimodal-Comp	-0.034	0.207	-0.162	0.871
	Monomodal vs. Bimodal-Comp	0.008	0.195	0.039	0.969
	Monomodal vs. Bimodal-NonComp	0.041	0.196	0.209	0.834

Table 23. Results of Rule Test for old trials and new trials.

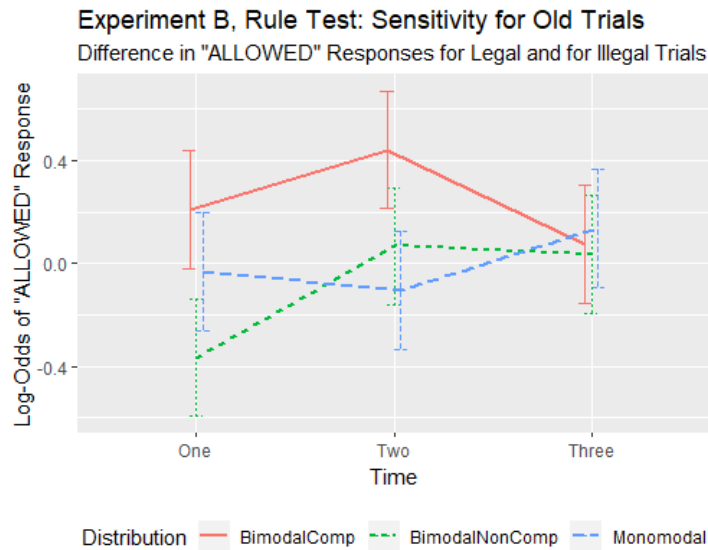


Figure 41. Results of Rule Test, old trials.

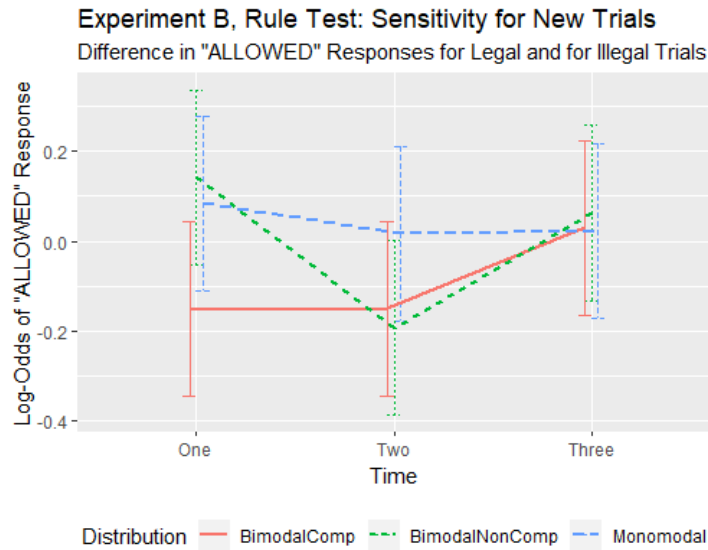


Figure 42. Results of Rule Test, new trials.

Since the Bimodal-Comp conditions were the only conditions which were exposed to the phonological rule that [ɛɑ] occurs after [i] and [ʂɑ] occurs after [u], one would only expect the groups exposed to Bimodal-Comp distributions to have greater sensitivity to the difference between *legal* and *illegal* trials if they had learned this phonological rule. Evidence of this is found for ExposureTime One for old trials, where the Bimodal-Comp group has significantly greater sensitivity to the difference between *legal* and *illegal* trials than the Bimodal-NonComp group (and numerically greater sensitivity than the Monomodal group), and also during ExposureTime Two for old trials (Figure 41), where the Bimodal-Comp group has significantly greater sensitivity to the difference between *legal* and *illegal* trials than the Monomodal group (and numerically greater sensitivity than the Bimodal-NonComp group). However, there are no significant differences between groups at any other ExposureTimes, even at ExposureTime Three. It is unclear why the Bimodal-Comp group shows evidence of learning the rule after 10 minutes of training (ExposureTime Two), but not after 15 minutes of training (ExposureTime Three). This may be an indication that the Rule Test used is not an accurate reflection of learners' knowledge of phonological rules they may have acquired, and/or that participants in ExposureTime Three were showing signs of fatigue. There

does not appear to be any evidence that the Bimodal-Comp group generalized any type of rule to the new generalization syllables that they had not heard during training (Figure 42).

## 6. Discussion

Experiment B trained participants on one of three distribution types (Bimodal-Comp, Bimodal-NonComp, and Monomodal distributions) for one of three ExposureTimes (5 minutes, 10 minutes, or 15 minutes) in order to examine the effect of environmental context over the time course of early category learning. It was predicted that either results would show evidence for 1) a two-stage model, in which the Bimodal-Comp and Bimodal-NonComp learning trajectories initially patterned together and showed greater initial sensitivity than the Monomodal learning trajectory in a first stage, and then in a second stage the Bimodal-Comp and Monomodal learning trajectories patterned together and showed less sensitivity than the Bimodal-NonComp learning trajectory; or 2) a one-stage model in which the Bimodal-Comp learning trajectory and Monomodal learning trajectory always patterned together. Neither of the predictions made by these models were clearly borne out in Experiment B, but this chapter concludes that results of this experiment are better explained by a one-stage model of allophonic acquisition than by a two-stage model. This is primarily because, numerically, the sensitivity in the Bimodal-Comp learning trajectory is always lower than that of the Bimodal-NonComp and Monomodal learning trajectories. This suggests the Bimodal-Comp group was never initially “tricked” into believing that there were two sound categories /ε/ and /ξ/. Therefore it does not appear this group ever solely took frequency distributions into account, disregarding environmental context.

The most unexpected learning trajectory was that of the learners trained on a Monomodal distribution. Numerically, learners trained on the Monomodal distribution showed the highest sensitivity to critical stimuli at ExposureTimes One and Two, and significantly higher sensitivity at ExposureTime Two compared to the Bimodal-Comp group. Although more research is needed, I believe this may suggest a period of prolonged uncertainty in the Monomodal group compared to the Bimodal-Comp group. That is, learners trained on a Bimodal-Comp distribution appear to have learned that critical items S<sub>1a</sub>-S<sub>4a</sub>

and S<sub>5a</sub>-S<sub>8a</sub> belong to a single phoneme quicker than the Monomodal group. This would be difficult to explain in a two-stage model in which bimodal learners initially learn that there are two phonetic categories, but I believe these results fit with Dillon and colleagues' one-stage model of phoneme acquisition. Dillon and colleagues start with the assumption that the learner knows the conditioning environments of each pair of Inuktitut allophones, but note that their model does not capture exactly *how* learners determine these environments. It follows that a learner must entertain a number of hypotheses regarding whether there are conditioning environments, and if so, what they are. Results from this study can be explained if a learner's hypothesis testing is modelled as a search specifically for multiple subsets with some conditioning environment relating subsets to one another, where learners only settle on the hypothesis that there is a single subset after all other natural conditioning environments have been tested. In this study, the Bimodal-Comp group quickly determined these conditioning environments and settled on the hypothesis that [ɕ] and [ʂ] are allophones of a single phoneme by ExposureTime Two, where [ɕ] occurs after [i] and [ʂ] occurs after [u]. On the other hand, the Monomodal group may still be entertaining and testing various hypotheses regarding possible conditioning environments, and so do not settle on the hypothesis that there is just one post-alveolar fricative phoneme as quickly as the Bimodal-Comp group does.

### 6.1. COMPARISON WITH EXPERIMENT A3

Because critical stimuli were identical to those used in Experiment A3 and the procedures between the two experiments were similar, this section provides a comparison of the current study's results and those in Experiment A3.

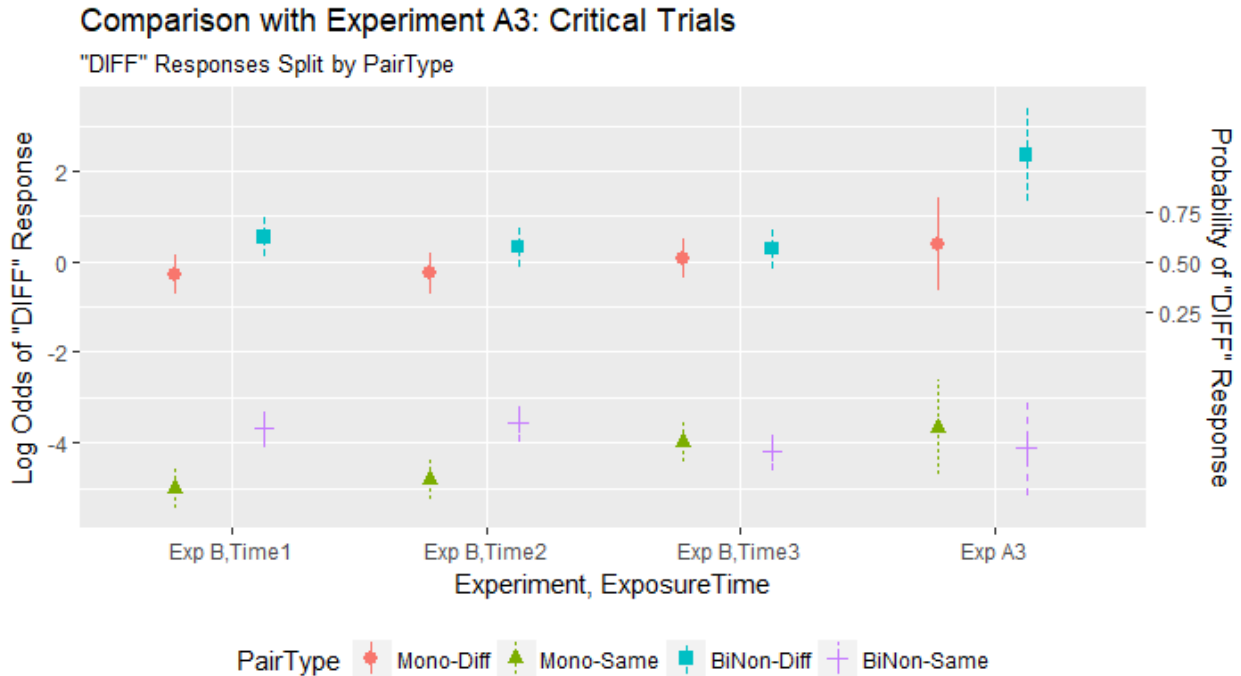


Figure 43. Log-odds of participants responding “different” for critical trials at all three ExposureTimes in Experiment B and in Experiment A3. Error bars indicate standard error.

Figure 43 shows the log-odds of a “different” response during the Phone Test of Experiment B and the only test phase in Experiment A3. Both of these test phases presented participants with pairs of syllables and asked participants to respond whether they believed these to be different syllables or repetitions of the same syllable. Experiment A3 found greater sensitivity for those trained on a bimodal distribution compared to those trained on a monomodal distribution. This can be seen in Figure 43 over the “Exp A3” label as greater sensitivity in the greater distance between the BiNon-Diff mean and the BiNon-Same mean, compared to the distance between the Mono-Diff mean and the Mono-Same mean. That is, those trained on a bimodal (non-complementary) distribution responded “different” more when presented with Different Pairs and “different” less when presented with Same Pairs, compared to those trained on a monomodal distribution. Participants in Experiment A3 were exposed to 192 critical syllables presented in isolation. Participants in ExposureTimes One, Two, and Three in Experiment B were exposed respectively to 96, 192, and 288 critical syllables presented in phrases. If one assumes the model presented in



the previous chapter, it appears that the presentation of critical syllables within phrases slowed the acquisition process since not even participants trained with 288 critical syllables showed a significant difference in sensitivity between the Bimodal-NonComp and Monomodal groups. This acquisition process may be slowed simply from the greater mental burden placed on learners from the need to process additional contextual syllables. Further ExposureTimes might have eventually led to a greater sensitivity in those trained on a Bimodal-NonComp distribution compared to those trained on a Monomodal distribution. However, more research is necessary to determine if this is the case. If further research were to be conducted, training should be split across more than one day, since this study finds evidence that the Bimodal-Comp group learned the phonological rule at ExposureTime Two, but not at ExposureTime Three, therefore possibly suggesting training fatigue for those in ExposureTime Three.

## 6.2. EARLIER EXPOSURE TIME

This study chose to train participants on one of three exposure times: one which lasted about 5 minutes, one which lasted about 10 minutes, and one which lasted about 15 minutes. I concluded that results were better explained by a one-stage model of allophony acquisition, since the Bimodal-Comp group was never “tricked” into initially patterning with the Bimodal-NonComp group. That is, this experiment was set up so that results would have to show a “hump” in sensitivity in the Bimodal-Comp group in order to determine there was evidence for a two-stage model (see the copy of Figure 35 below).

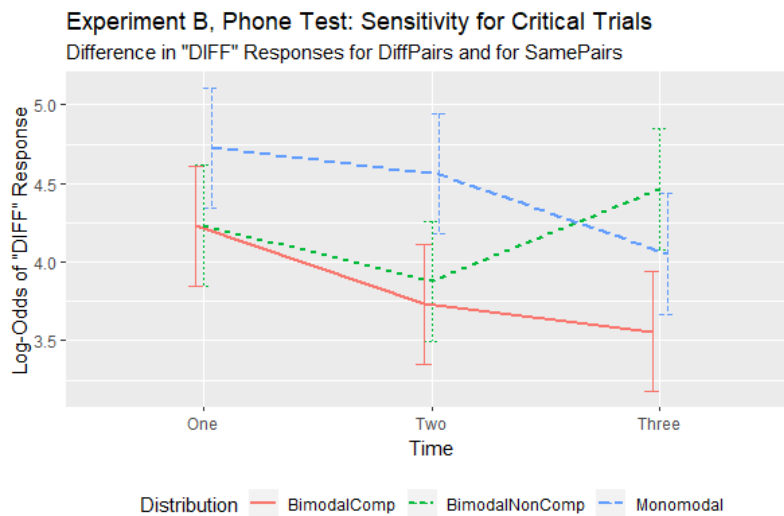
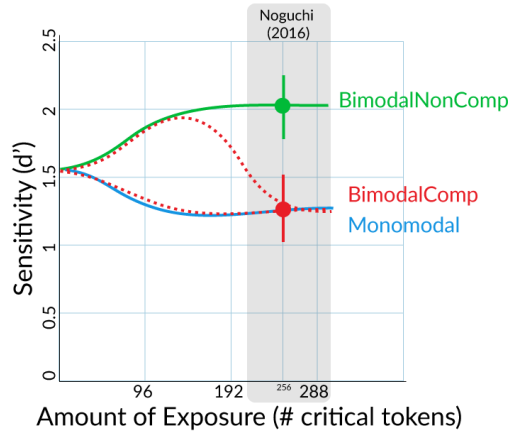


Figure 44. Comparison of the predicted change in sensitivity as amount of exposure increases (top; copy of Figure 35) and actual change in sensitivity as amount of exposure increases (bottom; copy of Figure 38).

One weakness of the current study was that neither prediction made by the one- or two-stage models was clearly borne out with significant differences in sensitivity between groups of learners. Even if we ignore this and draw conclusions from numerical results, another possible problem with this study is that ExposureTimes may not have been well-chosen. That is, it is possible that any initial “hump” in sensitivity in the Bimodal-Comp group occurred very early on in training, before the first ExposureTime tested (see Figure 45).

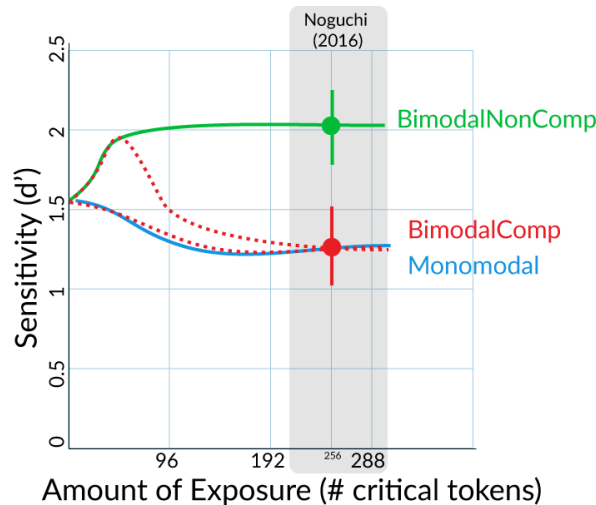


Figure 45. Hypothetical scenario illustrating an early “hump” in sensitivity for Bimodal-Comp group which could be missed by the first ExposureTime tested.

This possibility should be kept in mind, but I also do not believe this is a likely scenario, as this would require a significant amount of learning to have occurred before 5 minutes of training.

### 6.3. UNEXPECTED RESULTS OF FILLER STIMULI

As with the “A” Experiments, Experiment B also found unexpected results for filler test pairs. Specifically, the Bimodal-NonComp group showed significantly lower sensitivity than both the Bimodal-Comp and Monomodal groups for ExposureTime Two participants. Further research is necessary to determine whether this is an anomaly, or if something about the type and amount of exposure this group received would have some sort of effect on filler trials.

## 7. Conclusion

This experiment successfully replicates results from Noguchi (2016) on an online platform. Although results do not clearly support predictions of a one or a two-stage model of allophonic acquisition, results of this experiment are better explained by a one-stage model of allophony. At no point during the learning trajectory mapped in this study did the Bimodal-Comp group have higher sensitivities than the Monomodal group, and at ExposureTime 2 (which corresponded to hearing 192 critical stimuli and about 10

minutes of training) even exhibited significantly lower sensitivities than the Monomodal group. At ExposureTime 3 (which corresponded to hearing 288 critical stimuli and about 15 minutes of training), the Bimodal-Comp and Monomodal groups numerically appear to pattern together, with the Bimodal-Comp group having significantly lower sensitivities than the Bimodal-NonComp group. This is better explained by the one-stage model, which predicts that the Bimodal-Comp and Monomodal groups will pattern together throughout the time course of acquisition.

## **Chapter 5:**

### **A Gap in Both Bias and Sensitivity in Applications to Word Learning**

#### **1. Introduction**

In Chapter 3, it was argued that distributional learning occurs in two stages: a Bias Stage followed by a Sensitivity Stage. The current chapter explores the gap between learners' sound categories and use of those sound categories in lexical entries by examining whether learning from either the Bias Stage or the Sensitivity Stage extends to a word learning stage. Experiments C1 and C2, following assumptions based on results from Experiments A1 and A3 respectively, train and test participants over the course of three days, and find no evidence that either changes in bias or in sensitivity extend to a word learning task. Several proposals have been made to explain the gap in phone discrimination (i.e. sensitivity) and word learning (Pater et al., 2004; Werker and Curtin, 2005). The finding that there is also a gap in bias and word learning, necessitating an extension to these past proposals.

#### **2. Background**

The motivation for the "C" Experiments is two-fold. First, there is a well-documented gap in performance when it comes to discrimination and word learning (Pater et al., 2004; Stager and Werker, 1997; Werker et al., 2002). Since the "A" Experiments find evidence that distributional learning has two effects, one on response bias and one on sensitivity, the "C" Experiments were designed to test whether this gap in performance was found for both response bias and for sensitivity. The second observation motivating the "C" Experiments concerns the role of sleep in the consolidation of knowledge. Past studies have found that sleep is necessary for the integration of some knowledge with existing knowledge (Gaskell and Dumay, 2003; Leach and Samuel, 2007; Fenn et al., 2003). Following these observations, the "C" Experiments are conducted over the course of three days in order to determine whether a period of sleep is necessary to

overcome any gap between discriminatory abilities and word learning. This section begins with background on the gap in discrimination and word learning, followed by background on the possible role sleep may play.

## 2.1. A GAP BETWEEN PHONE DISCRIMINATION AND LEXICAL ACQUISITION

Numerous studies in both L1 and L2 acquisition point to a gap between the ability to discriminate phones, and the utilization of that discriminatory ability in distinguishing lexical items (L1: Pater et al., 2004; Stager and Werker, 1997; Werker et al., 2002; Hallé and de Boysson-Bardies, 1996; Kay-Raining Bird and Chapman, 1998; L2: Darcy et al., 2013; Daidone and Darcy, 2014; Weber and Cutler, 2004). Infants show language-specific discrimination of phones by the age of 12 months (Cheour et al., 1998; Kuhl et al., 2006; Kuhl et al., 1992; Mattock and Burnham, 2006; Mattock et al., 2008; Polka and Werker, 1994; Seidl et al., 2009; Werker and Tees, 1984; Best et al., 1995; Pegg and Werker, 1997; Tsao et al., 2006; Werker and Lalonde, 1988), but still confuse minimally-different words at 14 months (Werker et al., 2002; Werker et al., 1998; Pater et al., 2004). For example, Stager and Werker (1997) find that 14-month old infants trained on sound-meaning pairs (for example, Meaning A paired with Sound [bɪ] and Meaning B paired with Sound [dɪ]) failed to notice when sound-meaning pairs were switched with similar-sounding words (for example Meaning B paired with Sound [bɪ] rather than [dɪ]), despite being able to distinguish between [b] and [d] at the age of 8 months. This finding has been replicated with words conforming to English phonotactics (e.g. [bm] vs. [dn]), words differing in voicing (e.g. [bm] vs. [pʰɪn]), and words differing in both voicing and place of articulation (e.g. [pʰɪn] vs. [dn]) (Pater et al., 2004). Stager and Werker argue that infants use less phonetic detail when learning words than when, as they put it, “listening to syllables.” Follow-up studies find that this difficulty in learning minimally-differing words is particular to a certain age group and to a particular experimental design. Older infants (17 and 20 months) do notice a switch between the two minimally-differing sounds [bɪ] and [dɪ] in a word-learning task (Werker et al., 2002), suggesting that this failure to notice a switch occurs only for younger infants. Further studies find that the difficulty to learn minimally-differing words can be alleviated if variation is

introduced by way of multiple talkers (Rost and McMurray, 2009; 2010) or through a single talker (Galle et al., 2015), and also if conditions are conducive to word-referent mappings (Fennell and Waxman, 2010).

One could argue that the amount of exposure infants are given to a novel word is insufficient for the infant to form a phonetically-detailed lexical entry. How do infants fare with known words? Hallé and de Boysson-Bardies (1996) find evidence that 11-month olds treat high-frequency words that have been minimally altered still as high-frequency words, suggesting they do not notice phonetic detail even for familiar words. The 11-month old French infants showed preference to high-frequency words that had been altered into non-words (e.g. [kato] for *gâteau*) over low-frequency words (e.g. [byzar] *busard*). Further, the infants did not show any preference between high-frequency unaltered words (e.g. [gato] *gâteau*) and high-frequency altered words (e.g. [kato] for *gâteau*). Results suggest that infants still process altered high-frequency words as being high-frequency words, failing to utilize phonetic detail to distinguish even familiar words. Hallé and de Boysson-Bardies suggest a distinction between a “lexical mode,” in which infants are focused on word recognition and ignore phonetic details, and a “neutral mode,” in which infants are not concerned with word recognition or comprehension and instead are focused on the details of the pronunciation.

Hallé and de Boysson-Bardies found evidence for a lexical frequency preference, where frequency was determined by overall wordform rather than requiring the label to be an exact match to an existing label. That is, infants appeared not to have noticed the mispronunciation. While Hallé and de Boysson-Bardies’s study did not require a label-referent mapping, “mispronunciation sensitivity” studies aim to determine when a mispronunciation is great enough to interfere with this label-referent mapping. That is, is the minimally differing label treated by the language learner as being an equally good label of the referent as the correct label? Swingley and Aslin (2000) test this by presenting 18-23 month olds with two pictures. Children heard sentences containing a label for one of the pictures that was either pronounced correctly (*baby*) or mispronounced (*vaby*). For both the correctly pronounced and the

mispronounced label, children correctly identified the referent, but did so more slowly when the label had been mispronounced. A recent meta-analysis finds no developmental change in mispronunciation sensitivity from 6-30 months (Von Holzen and Bergmann, 2018).

That listeners use less phonetic detail in lexical tasks than in discriminatory tasks is mirrored in L2 studies as well. For example, Spanish contains an alveolar tap /ɾ/ and an alveolar trill /r/. In intervocalic positions, switching one phone with the other causes a change in meaning (e.g. [pero] ‘but’ compared to [perɾo] ‘dog’). In a lexical decision task, English-speaking L2 learners of Spanish fail to identify Spanish words with intervocalic /ɾ/ pronounced as [r] or intervocalic /r/ pronounced as [ɾ] (e.g. *quiero* /kiero/ pronounced as [kiero]) as being non-words in Spanish, despite being able to successfully classify minimal pairs as containing intervocalic [ɾ] or [r] in an ABX task and therefore being able to discriminate between the two sounds (Daidone and Darcy, 2014).

In artificial language learning studies, Hayes-Harb (2007) examines this gap in a replication study of Maye and Gerken (2000), followed by a word-learning component. She first trained one group of participants on a bimodal distribution of phones (from voiceless unaspirated stop [q̥] to prevoiced stop [q̚]), and another group on a monomodal distribution of phones. As predicted by Maye and Gerken (2000), learners in the monomodal group were more likely to say that the ends of the continuum (e.g. [q̥a] vs. [q̚a]) were the “same.” Subsequently, participants from both groups were trained on lexical items. For example, participants heard [q̚ant] and were told that the meaning of this word was ‘boot.’ Following this word-training, participants were tested on whether participants in the bimodal group had encoded the two ends of the continuum [q̥] and [q̚] differently in the words they were trained on more than the monomodal group had. The author did so by testing participants on the opposite ends of the continuum: if they had been trained that [q̚ant] was the word for ‘boot’, they were asked whether [q̥ant] was a mispronunciation. She did not find any difference between the bimodal group and the monomodal group in this word-learning task.



On the other hand, Perfors and Dunbar (2010) do find evidence that distributional learning aids a following word learning task. In their study, participants are placed in either a bimodal or control group (there was no monomodal condition). Participants in the bimodal group heard 912 critical tokens with onsets ranging between [g] and [g̃] and no fillers, while participants in the control group heard 912 tokens beginning with [d] and [tʰ], but no critical tokens. Following this intensely focused training task with no fillers, participants were tested on discrimination of [g] and [g̃] in an ABX task. In a following word learning task, participants were presented with minimal pairs each paired with some picture indicating the word's meaning. For example, learners were trained that one image was a [g̃]ipur, while another was a [g]ipur. Perfors and Dunbar find that participants in the bimodal group (1) responded correctly on more critical trials in the ABX discrimination task than the control group, and (2) responded correctly to the word learning phase more than the control group.

Although Perfors and Dunbar's study appears to find evidence that learners trained on a bimodal distribution can extend their distributional learning to a following word learning task, there are a few weaknesses which could be improved. First, their study does not test a monomodally-trained group of participants. Therefore, it is unclear from their study whether the bimodal learners are simply better at the task because they receive more exposure to critical stimuli, or whether their performance really was due to distributional learning. Second, Perfors and Dunbar test learners with an ABX discrimination task and use minimal pairs in their word learning task. As discussed in Chapter 2, an ABX task already implicitly tells participants that there exist (at least) two critical categories and, according to the model put forward in this dissertation, places learners at the second stage of distributional learning. The use of minimal pairs in the word learning task has the same effect, since learners are being clearly shown that there are two similar sounds with different semantic referents. Third, Perfors and Dunbar's study did not include any filler items. Although not necessarily an issue, it is possible that the lack of fillers caused participants to use more explicit methods of learning, and to form specific hypotheses. This dissertation does not make specific claims regarding the implicit or explicit nature of distributional learning, but the lack of fillers in

Perfors and Dunbar (2010) is still important to keep in mind. For these reasons, the current study will follow the procedure of Hayes-Harb (2007), which I believe more closely imitates natural language acquisition, with the exception that the current study will increase exposure to take place over the course of three days.

### 2.1.1 *Past proposals*

How has this gap between phone discrimination and word learning been explained by past researchers? Although a number of proposals have been made, this section will focus on two: one by Werker and Curtin (2005), and one by Pater et al. (2004).

Werker and Curtin (2005) describe an all-encompassing framework, PRIMIR (Processing Rich Information from Multidimensional Interactive Representations), to explain early linguistic development. This model is made up of three planes: the General Perceptual plane, which contains phonetic information; the Word Form plane, which contains meaningless extracted units; and the Phonemic plane, which is made up of language-specific phonetic categories. In addition, there are three filters: initial biases, task demands, and developmental level. These filters determine where the learner directs his or her attention in regards to each of the three planes (see Figure 46) for a given task.

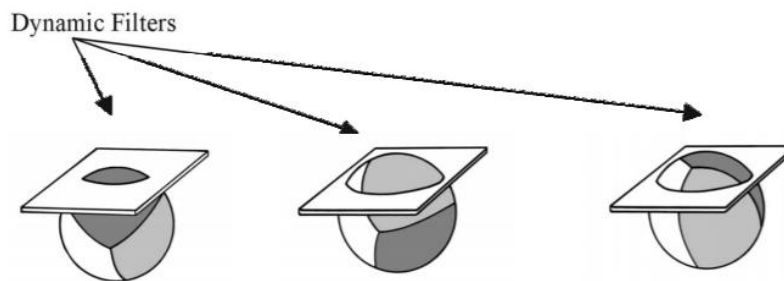


Figure 46. Effect of filters (shown by the rectangle) on attention to each of the three planes (shown by different shades on the ball). Figure from Werker and Curtin (2005).

For a simple discrimination task, a participant’s “task demands” filter may be shifted so that only the General Perceptual plane is engaged (for example, the leftmost schematic in Figure 46, where the darkest

shade indicates the General Perceptual plane). For tasks which require matching word forms with concepts, the task demands filter may shift so that the participant's attention is balanced evenly across all three planes, with the end result that less of the General Perceptual plane is engaged than it was for the discrimination task. In this model, the learner's overall cognitive capacity (in the form of the "developmental level" filter), as well as his or her attention levels and where attention is directed (in the form of the "task demands" filter), plays a role in how much phonetic detail is being utilized in a given task. The gap we see between discrimination and the application of discriminatory ability to a word learning task is the result of attention being shifted away from the General Perceptual plane.

Pater et al. (2004) also rely on explaining this gap with attention levels, although they formalize their model in a different way. Their model is framed within Optimality Theory. In their model, faithfulness constraints are demoted with respect to markedness constraints when a listener's attention or cognitive resources are diminished. This follows proposals such as Boersma (1998) and Davidson et al. (2004). Lower faithfulness constraints result in a simplification of the perceived structure, resulting in neutralization of a contrast. Therefore, for Pater et al., the gap between phone discrimination and word learning is the result of fewer available resources due to the greater cognitive complexity which a word learning task demands compared to a discrimination task, which in turn lowers faithfulness constraints.

While both of these proposals center around the role of attention, neither explicitly make any claims regarding response bias. This study will argue that bias as well as sensitivity should also be included as a factor affected by attention levels.

## 2.2. A POSSIBLE ROLE OF SLEEP

One possible reason Hayes-Harb (2007) failed to find evidence for distributional learning extending to a word learning phase is that participants were not exposed to enough tokens. Another possible reason is that the exposure period took place over a single day, and there is also reason to think that participants need a period of sleep in order to integrate learned information with existing information (Gaskell and Dumay, 2003; Leach and Samuel, 2007; Fenn et al., 2003). Gaskell and Dumay (2003) find that newly-

learned words only exhibit lexical competition effects after a period of sleep, arguing that there are two stages of word learning: “lexical engagement” (also “phonological learning”) and “lexical consolidation” (Leach and Samuel, 2007; Gaskell and Dumay, 2003). Lexical engagement is an initial stage of word learning in which the phonological form is quickly acquired, while the second stage is slower, and refers to the integration of this form with existing information. Therefore it is possible that Hayes-Harb did not find evidence that participants extended newly-learned phonetic categories to a word-learning task simply because participants needed to go through a period of sleep.

### 3. Research Question and Summary

This study seeks to further explore the gap between discriminatory abilities and phonological knowledge in word learning by testing whether this gap exists for both response bias and sensitivity. Two research questions asked in this chapter are as follows:

- (1) Is there evidence that changes in **response bias** caused by distributional learning extend to a word learning task, either before or after a period of sleep?
- (2) Is there evidence that changes in **sensitivity** caused by distributional learning extend to a word learning task, either before or after a period of sleep?

In addition, this study will test the following hypothesis:

- (3) Participants in Hayes-Harb (2007) simply needed a period of sleep to integrate newly-learned phonetic category information with word-related information.

This study finds no evidence that changes in bias or sensitivity as gained from distributional learning extend to a word learning task, even over the course of three days of training.

Two experiments are presented in this chapter: Experiments C1 and C2. Experiments C1 and C2 are based on the assumption that distributional learning has had some sort of effect on learners. Specifically, Experiment C1 assumes that, by the end of training, bimodal learners have a greater bias towards a “different” response compared to monomodal learners. This assumption is based on results of Experiment

A1, which finds that after exposure to 192 critical stimuli ranging between [g] and [ɟ] and 192 fillers, bimodally-trained participants exhibit greater bias towards a “different” response than monomodally-trained participants. Experiment C1 will be identical to Experiment A1 up through the training phase (see Figure 47).

Similarly, Experiment C2 assumes that, by the end of training, bimodal learners have greater sensitivities to the endpoints of the critical stimulus continuum. This assumption will be based on results of Experiment A3, which used the critical stimuli [ɛɑ] – [ʂɑ]. Experiment C2 will be identical to Experiment A3 up through the training phase. Figure 47 gives a comparison of the procedure for each experiment. The boxed portions are completely identical between Experiment A1 and Experiment C1, and between Experiment A3 and Experiment C2. Because of this, this chapter will assume that the results of Experiments A1 and A3 are indicative of participants’ behaviors when they begin the word learning phase in Experiments C1 and C2, respectively.

Experiments A1, A3	Experiments C1, C2		
	DAY 1	DAY 2	DAY 3
	Login	Login	Login
	---	Sleep questionnaire	Sleep questionnaire
Sound check	Sound check	Sound check	Sound check
English practice word test	English practice word test	English practice word test	English practice word test
Training	Training	Training	Training
	Word learning	Word learning	Word learning
Phone test	Word test	Word test	Word test
Questionnaire	---	---	Questionnaire

Figure 47. Comparison of Experiments A1/A3 (left) with Experiments C1/C2 (right). The boxed portions in Experiment A1 is completely identical to the portion in Experiment C1, and the boxed portion in Experiment A3 is completely identical to the portion in Experiment C2.

Experiments C1 and C2 primarily differ in the stimuli being used and, because of that, what stage of distributional learning learners are assumed to be in by the time they enter the word learning phase.

Specifically, Experiment C1 assumes that learners have entered the bias stage by the end of the first day’s training phase, and Experiment C2 assumes that learners have entered the sensitivity stage by the end of the first day’s. A comparison of Experiments C1 and C2 is shown in Figure 48.

	Experiment C1	Experiment C2
Critical stimuli	[g] – [g]	[ç] – [ʂ]
Assumption	Distributional learning results in greater bias towards a “different” response in the bimodal group compared to the monomodal group, for these particular stimuli in the training period given	Distributional learning results in greater sensitivities in the bimodal group compared to the monomodal group, for these stimuli in the training period given
Assumption based on...	Results of Experiment A1	Results of Experiment A3
Preview of results	No evidence that changes in response bias from distributional learning extend to a word-learning task	No evidence that changes in sensitivities from distributional learning extend to a word-learning task

Figure 48. Comparison of Experiments C1 and C2.

**4. Experiment C1**

Experiment C1 tests Research Questions 1 and 3. Specifically, Experiment C1 tests for evidence of a gap in response bias between a discrimination task and a word-learning task, and, if there is such evidence, it tests whether sleep is necessary to overcome this gap.

4.1. METHODOLOGY

Experiment C1 closely follows the methodology of Hayes-Harb (2007). The main differences between Hayes-Harb (2007) and the current experiment are that (1) this study only trains participants on a bimodal and monomodal distribution, (2) and participants participate in Train-WordLearning-Test phases repeated each day over the course of three consecutive days in order to test the effect of increased exposure.

4.2. STIMULI

Stimuli consisted of four types of syllables: no-coda critical syllables, no-coda filler syllables, coda critical syllables, and coda filler syllables. No-coda filler syllables were identical to the filler syllables used in Experiment A1. No-coda critical syllables were also identical to those used in the Experiment A1 and consisted of onsets drawn from an 8-point continuum ranging between [g] and [ç], followed by one of three nuclei, [a æ ɪ]. Details of the creation of this continuum can be found in Chapter 3. Examples of G<sub>1</sub>a and G<sub>8</sub>a are copied from Chapter 3, in Figure 23.

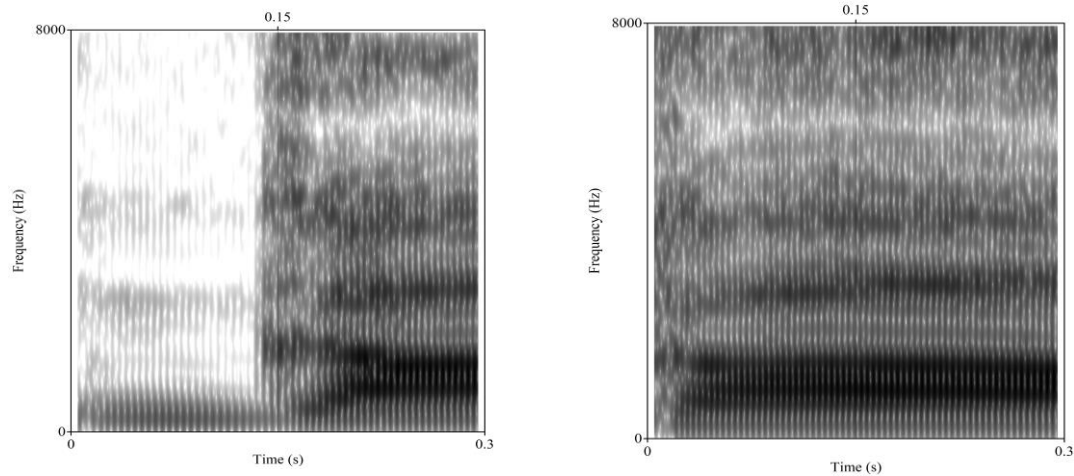


Figure 49. Initial 300 ms of spectrograms of  $G_{1\alpha}$  (left) and  $G_{8\alpha}$  (right) for critical stimuli.

Coda critical syllables were created by splicing codas to shortened no-coda critical syllables. To obtain codas, the following six syllables were recorded: [gamp], [gasp], [gæŋ], [gævz], [gɪk], [gɪf]. The codas for each of these syllables were spliced out, with cuts being made at positive-going zero crossings. The ends of no-coda critical syllables were removed (with cuts also being made at positive-going zero crossings). The lengths removed were based on trial and error, and ranged from 60-70 ms for each syllable. I tried to ensure that the amount removed from the ends of each syllable were consistent within each of the three continua (one for each nucleus type [a æ ɪ]). No shortened no-coda critical syllable sounded irregular to my ears when played side by side with other members of the same continuum. Before concatenating shortened no-coda critical syllables and codas, the intensity of codas were either raised or lowered. The amount and direction that coda intensities were altered so that they would sound most natural when spliced together with the shortened no-coda critical syllables was determined through trial and error. After concatenating shortened no-coda critical syllables and (intensity-modified) codas, pitch blips at the concatenation point were smoothed out if necessary. This was done using Praat's pitch tier editor. If a pitch modulation was detectable at the splice point, 3-4 points from the pitch tier at the point of concatenation were removed. After manipulations were completed, all 8 members of a word "series" (e.g.  $G_{1-8}$  amp) were played side by side to ensure that all members sounded uniform.

Coda filler syllables were simply recorded and were not created through concatenation. Two tokens of each of the following filler syllables were recorded: [fas], [fæs], [təb], [tejb], [mæfs], [næfs], [sem], [zɛm].

#### 4.3. PROCEDURE

Participants were again recruited through Mechanical Turk. This experiment took place over the course of three days. Experiment C1 consisted of 8 phases, only some of which were presented on each day of the experiment. A summary of the phases presented on each day is shown in Figure 50.

Day 1	Day 2	Day 3
Login	Login	Login
---	Sleep questionnaire	Sleep questionnaire
Sound check	Sound check	Sound check
English practice word test	English practice word test	English practice word test
Training	Training	Training
Word learning	Word learning	Word learning
Word test	Word test	Word test
---	---	Questionnaire

Figure 50. Summary of procedure on each day. The double-boxed portion indicates a procedure identical to that found in Experiment A1 before the test phase.

Each day began with a **Login**, which prompted users to enter their Mechanical Turk ID. This ensured that the same participant was logging in each day. On Days 2 and 3, this was followed by a **Sleep Questionnaire**, which asked participants how many hours they had slept the previous night, and how they would rate the quality of their sleep. Following this was a **Sound Check**, identical to that used in the “A” Experiments. Again, participants heard one or two 50 Hz tones in order to ensure that participants were wearing headphones. Participants were instructed to press the “1” or the “2” key to indicate the number of tones they had heard.

During each block of **Training**, participants heard 16 tokens drawn from each of the three critical no-coda continua, for a total of 48 critical no-coda syllables per block. Additionally, they heard two repetitions of three different tokens of filler no-coda syllable: [fa fæ tɛ tej mə nə sɛ zɛ], for a total of 48



fillers per block. Training on each day consisted of 4 blocks, for a total of 192 critical syllables and 192 filler syllables per day.

Following training, participants were directed to a **Word-Learning phase**. The word-learning phase presented participants with sound-meaning pairs that made up words in this artificial language. Sounds consisted of critical and filler coda syllables. For critical items, onsets of each “word” consisted of either  $G_{1-4}$  or  $G_{5-8}$ . Which half of the continuum the word began with was determined by the learner’s subcondition. For instance, the sound corresponding to the meaning ‘apple’ in this experiment was *Gamp*, where “G” consisted of the first four continuum points  $G_{1-4}$  if participants were in subcondition A, and  $G_{5-8}$  if in subcondition B. Each participant heard all 4 tokens (corresponding to the 4 continuum points of each half of the critical continuum) of each critical word once during this word learning phase. Therefore participants saw a picture of each critical word’s meaning 4 times, and heard each of the word’s 4 auditory tokens once, on each of the 3 days of this experiment. For filler trials, participants only heard one repetition of two auditory tokens paired with some meaning (rather than 4, in order to keep this phase short). Again, the sound paired with some meaning depended on the participant’s subcondition. See Figure 51 for a list of sound-meaning pairs presented to each subcondition.

<b>Subcondition A</b>	<b>Subcondition B</b>	<b>Meaning</b>
$G_{1-4}$ amp	$G_{5-8}$ amp	apple
$G_{5-8}$ asp	$G_{1-4}$ asp	fork
$G_{1-4}$ æŋ	$G_{5-8}$ æŋ	chair
$G_{5-8}$ ævz	$G_{1-4}$ ævz	boot
$G_{1-4}$ ɪk	$G_{5-8}$ ɪk	elephant
$G_{5-8}$ ɪf	$G_{1-4}$ ɪf	lamp
fas <sub>1</sub> , fas <sub>2</sub>	fæs <sub>1</sub> , fæs <sub>2</sub>	motorcycle
tɛb <sub>1</sub> , tɛb <sub>2</sub>	tejb <sub>1</sub> , tejb <sub>2</sub>	onion
mæfs <sub>1</sub> , mæfs <sub>2</sub>	næfs <sub>1</sub> , næfs <sub>2</sub>	horse
sem <sub>1</sub> , sem <sub>2</sub>	zɛm <sub>1</sub> , zɛm <sub>2</sub>	trumpet

Figure 51. Sound-meaning pairs presented during the word learning phase.

For each trial of the word learning phase, the auditory stimulus was played and the picture paired with that sound was presented on the screen for 2000 ms, before the next trial began. Participants were instructed to simply try to memorize the words presented in this phase, and were reminded not to write anything down.

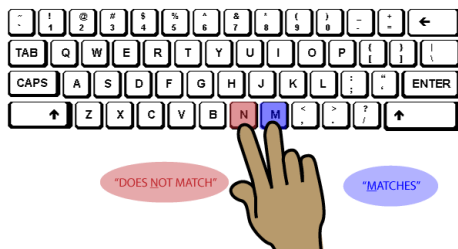
Following the word learning phase, participants were directed to a test phase. Before the test phase, participants were given the following directions:

*Great work! This next phase is the TESTING phase of the experiment, and will be similar to the practice test you did earlier in English. **This time it will ask you about the foreign language that you just heard.***

*Like before, please place one finger over the "M" and another finger over the "N" key on your keyboard.*

*Like before, if you think the word you hear and the image you see **MATCH**, press the "M" key.*

*If you think the word you hear and the image you see **do NOT match**, press the "N" key.*



*Please use what you have learned about the language during the LISTENING phase. If you are not sure, please make your best guess and move on to the next pair.*

In the word test phase, participants were presented with randomized trials of sound-meaning pairs. In each trial, a picture was shown on the screen, and participants heard a sound. For each trial, sound-meaning pairs were either matched (MatchedPair) or mismatched (MismatchedPair). MatchedPair trials presented the same sound-meaning pair participants had been trained on during the word learning phase. MismatchedPair trials presented the same picture participants had seen, but the auditory stimulus did not match what participants had heard during the word learning phase. For critical Mismatched trials, participants were presented with an auditory stimulus whose onset was drawn from the opposite end of the continuum as those they had been exposed to during the word learning phase. For example, a participant

in subcondition A was heard  $G_{1-amp}$  paired with the meaning ‘apple’ during the word learning phase. During testing, MatchedPair trials consisted of the audio stimulus  $G_{1amp}$  paired with a picture of an apple, and MismatchedPair trials consisted of the audio  $G_{8amp}$  paired again with a picture of an apple. Only continuum points  $G_1$  and  $G_8$  were presented during testing.

For filler trials, MatchedPairs again consisted of a sound-meaning pair that participants had been exposed to during the word learning phase. Sounds in MismatchedPairs differed from audio participants had been trained on by one segment (e.g. [fas] / [fæs]).

Meaning	Matched (for Subcond. A) Mismatched (for Subcond. B)	Mismatched (for Subcond. A) Matched (for Subcond. B)
apple	$G_1 amp$	$G_8 amp$
fork	$G_8 asp$	$G_4 asp$
chair	$G_1 æŋ$	$G_8 æŋ$
boot	$G_8 ævz$	$G_4 ævz$
elephant	$G_1 .lk$	$G_8 .lk$
lamp	$G_8 .lʃ$	$G_1 .lʃ$
motorcycle	fas	fæs
onion	tɛb	tejb
horse	mæfs	næfs
trumpet	sɛm	zɛm

Figure 52. Summary of MatchedPairs and MismatchedPairs.

A brief summary of each phase presented in this experiment is shown in Figure 53.

<b>Phase</b>	<b>Abbreviated directions</b>	<b>Example</b>
Login		
Sleep questionnaire (only on Days 2 and 3)	<i>How well did you sleep? How much sleep did you get?</i>	
Sound check	<i>Press the '1' if you hear 1 beep, the '2' if you hear 2</i>	
English practice word test	<i>Press 'm' if the sound <u>m</u>atches the picture you see, 'n' if they do <u>n</u>ot match</i>	<i>disk</i> (see picture of a desk)
Training	<i>Listen carefully</i>	<i>G<sub>3a</sub></i> <i>ma</i>
Word learning	<i>Try to memorize these words</i>	<i>G<sub>1-amp</sub></i> (see picture of apple)
Test	<i>Press 'm' if the sound <u>m</u>atches the picture you see, 'n' if they do <u>n</u>ot match</i>	<i>G<sub>8amp</sub></i> (see picture of apple)
Questionnaire (only on Day 3)		

Figure 53. Summary of each phase in Experiment C1.

A total of 69 participants participated on the first day. 52 of these participants took all three days of the experiment. Because the questionnaire took place on the third day, only participants who had participated all three days were considered for analysis. Participants were excluded from analysis if they reported in their questionnaire answers that they: (1) had a history of a speech or hearing disorder (3 participants excluded for this reason); (2) were not native speakers of English (1 participant excluded for this reason); (3) wrote down words (1 participant excluded for this reason<sup>28</sup>).

Participants were also excluded if they: (1) received a combined score across all three days of less than 15/18 on the sound check (11 participants excluded for this); or (2) a combined score across all three days of less than 15/24 on the English practice test (0 excluded for this). This left a total of 37 participants (18 bimodal, 19 monomodal), with some participants excluded for multiple reasons.

---

<sup>28</sup> This participant reported writing down the word for *trumpet*, saying they had thought they would not remember it. They also reported that they did not end up needing to refer back to it.

#### 4.4. ANALYSIS

##### 4.4.1 Model to be used in analysis: Bias and Sensitivity Tests

The regression formula described below models one dependent variable, **Response**, with three fixed effects: (1) **Distribution**, a between-subject, within-item factor consisting of two levels {*bimodal*, *monomodal*}, (2) **PairType**, a within-subject, between-item factor consisting of two levels {*matchedPair*, *misMatchedPair*}, and (3) **Day**, a within-subject, within-item factor consisting of three levels {*day1*, *day2*, *day3*}. This was done separately for *critical* and *filler* trials. The dependent variable **Response** consists of two levels, *m* and *n*, where *m* corresponds to a participant response that the sound-picture pair “**m**atched” during the Test phase, and where *n* corresponds to a participant response that the sound-picture pair “did **n**ot match” during the Test phase. Random effects for **Subject** and **Item** were also included in the model described below. All variables are summarized in Table 9.

Variable type	Effect type	Factor name	Factor type	Level names	Description
Independent variables	Fixed effects	Distribution	Between-subject, within-item	<i>bimodal</i> <i>monomodal</i>	Distribution type received by the participant during Train phase
		PairType	Within-subject, between-item	<i>matchedPair</i> <i>misMatchedPair</i>	Type of pair presented during Test phase (MatchedPair or MismatchedPair)
		Day	Within-subject, within-item	<i>day1</i> <i>day2</i> <i>day3</i>	Day session
	Random effects	Subject			Each individual participant (coded by ID)
		Item			Each individual item presented during the Test phase
Dependent variables		Response		<i>m</i> <i>n</i>	Response given by participant

Table 24. Variables used in regression analysis for Experiment C1, testing for bias and sensitivity.

As with previous chapters, all statistical tests were completed in R (R Core Team, 2014), using the `glmer` function from the `lme4` package (Bates et al. 2015) to fit a generalized linear mixed-effects model (GLMM) with a logistic link function (“mixed logit model”). Significance was set at a level of  $p <$

0.05. Two separate regressions were conducted to compare the effect of Distribution and PairType on Response: one for *critical* items, and one for *filler* items. Continuing the procedure for analysis used in previous chapters, this chapter follows suggestions made by Clark (1973) for variables to include as random effects, and suggestions made by Barr et al. (2013) and Barr (2013) for the random effects structure of the formula fitted in the regression by including random slopes for the highest-order combination of within-unit factors (PairType\*Day for Subject; Distribution\*Day for Item)<sup>29</sup>. The formula initially fitted to the regression is shown in (1).

$$(1) \quad \text{Response} \sim \text{Distribution*PairType*Day} + (1+\text{PairType*Day}|\text{Subject}) + (1+\text{Distribution*Day}|\text{Item})$$

The formula in (1) failed to converge, so the random effects structure was simplified to that shown in the formula in (2), and an Adaptive Gauss-Hermite Quadrature algorithm was used (by setting nAGQ to 0 in R) rather than the default Laplace Approximation algorithm.

$$(2) \quad \text{Response} \sim \text{Distribution*PairType*Day} + (1+\text{PairType}|\text{Subject}) + (1+\text{Distribution}|\text{Item})$$

#### 4.4.2 Model interpretation

Experiment C1 is primarily interested in testing response bias; however, for completeness, both **bias** and **sensitivity** will be tested for here. As before, results of the regression are interpreted as follows: an interaction of Distribution and PairType will be interpreted as a difference between the bimodal and monomodal conditions in **sensitivity** (referred to here as the Sensitivity Test). A significant main effect of Distribution *without* a significant interaction between Distribution and PairType is interpreted as evidence for a difference in **bias** between the bimodal and monomodal conditions (referred to here as the Bias Test). As in previous chapters, a main effect of Distribution with a significant interaction will not be considered interpretable in this analysis (see Chapter 2 for a discussion).

---

<sup>29</sup> See Chapter 3 for details regarding the reasoning behind the random effects structure used.

#### 4.4.3 Model to be used in analysis: Sleep Tests

This study also tests for an effect of amount of sleep. The regression formula used to test for an effect of sleep models **Response**, with three fixed effects: (1) **Distribution** {*bimodal*, *monomodal*}, (2) **PairType**{*matchedPair*, *misMatchedPair*}, and (3) **SleepAmount**, factor consisting of two levels {*less*, *more*}, where “less” applied to subject responses if they reported having slept less than 8 hours the night before, and “more” applied to subject responses if they reported having slept 8 or more hours the night before. This will be done separately for *critical* and *filler* trials. The dependent variable **Response** consists of two levels, *m* and *n*, where *m* corresponds to a participant response that the sound-picture pair “matched” during the Test phase, and where *n* corresponds to a participant response that the sound-picture pair “did not match” during the Test phase. Random effects for **Subject** and **Item** were also included in the model described below. To do this, the formula in (3) was used.

$$(3) \quad (\text{Response} \sim \text{Condition} * \text{PairType} * \text{SleepAmount} + (1 + \text{PairType} | \text{Subject}) + (1 + \text{Condition} | \text{Item}))$$

### 4.5. RESULTS

#### 4.5.1 Bias and Sensitivity Tests

Results of the regression fitted to formula (2) are shown in Table 25. Figure 54 and Figure 55 display results of Experiment C1 split by PairType.

Predictor	Coefficient	SE	Wald Z	p
<b>CRITICAL TRIALS</b>				
(Intercept)	-2.033	0.377	-5.397	<0.001 ***
Distribution= <i>monomodal</i>	0.262	0.484	0.542	0.588
PairType= <i>misMatchedPair</i>	-0.527	0.470	-1.12	0.263
Day= <i>two</i>	-1.351	0.608	-2.222	0.026
Day= <i>three</i>	-1.650	0.732	-2.254	0.024
Interaction= <i>monomodal &amp; misMatchedPair</i>	0.414	0.616	0.672	0.502
Interaction= <i>monomodal &amp; two</i>	0.817	0.763	1.072	0.284
Interaction= <i>monomodal &amp; three</i>	-0.031	0.976	-0.031	0.975
Interaction= <i>misMatchedPair &amp; two</i>	0.613	0.790	0.777	0.437
Interaction= <i>misMatchedPair &amp; three</i>	-1.410	1.235	-1.141	0.254
Interaction= <i>monomodal &amp; misMatchedPair &amp; two</i>	-1.088	0.990	-1.099	0.272
Interaction= <i>monomodal &amp; misMatchedPair &amp; three</i>	2.229	1.399	1.593	0.111
<b>FILLER TRIALS</b>				
(Intercept)	-2.544	0.609	-4.177	<0.001 ***
Distribution= <i>monomodal</i>	0.144	0.753	0.191	0.849
PairType= <i>misMatchedPair</i>	2.700	0.612	4.414	<0.001 ***
Day= <i>two</i>	-0.323	0.649	-0.498	0.618
Day= <i>three</i>	-0.510	0.724	-0.704	0.481
Interaction= <i>monomodal &amp; misMatchedPair</i>	0.037	0.845	0.044	0.965
Interaction= <i>monomodal &amp; two</i>	0.240	0.874	0.275	0.784
Interaction= <i>monomodal &amp; three</i>	0.726	0.931	0.779	0.436
Interaction= <i>misMatchedPair &amp; two</i>	0.939	0.744	1.261	0.207
Interaction= <i>misMatchedPair &amp; three</i>	1.322	0.785	1.682	0.093
Interaction= <i>monomodal &amp; misMatchedPair &amp; two</i>	-0.833	1.003	-0.831	0.406
Interaction= <i>monomodal &amp; misMatchedPair &amp; three</i>	-0.851	1.032	-0.825	0.410

Table 25. Results of GLMM for Experiment C1.

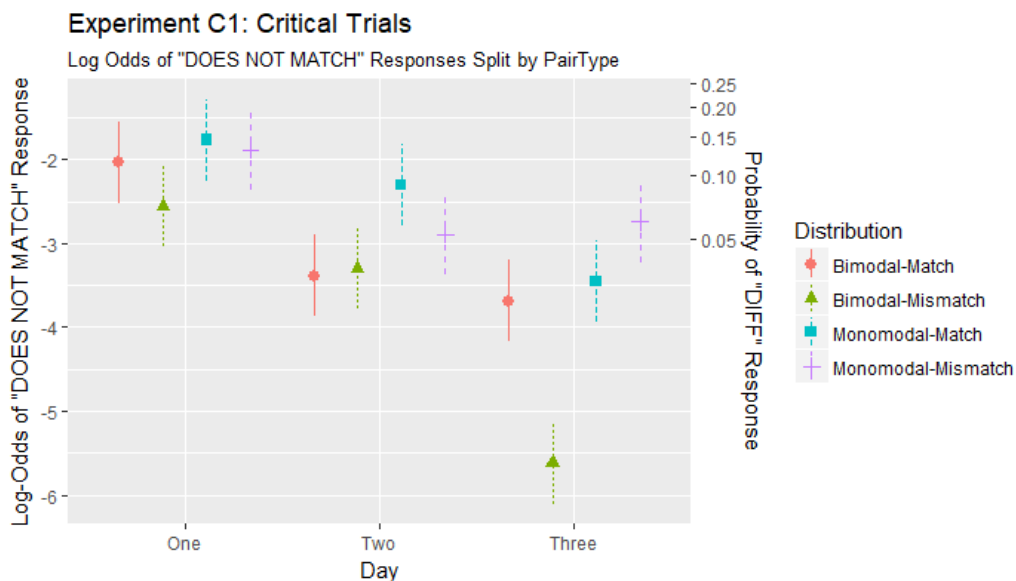


Figure 54. Results from Experiment C1 split by PairType, critical trials.



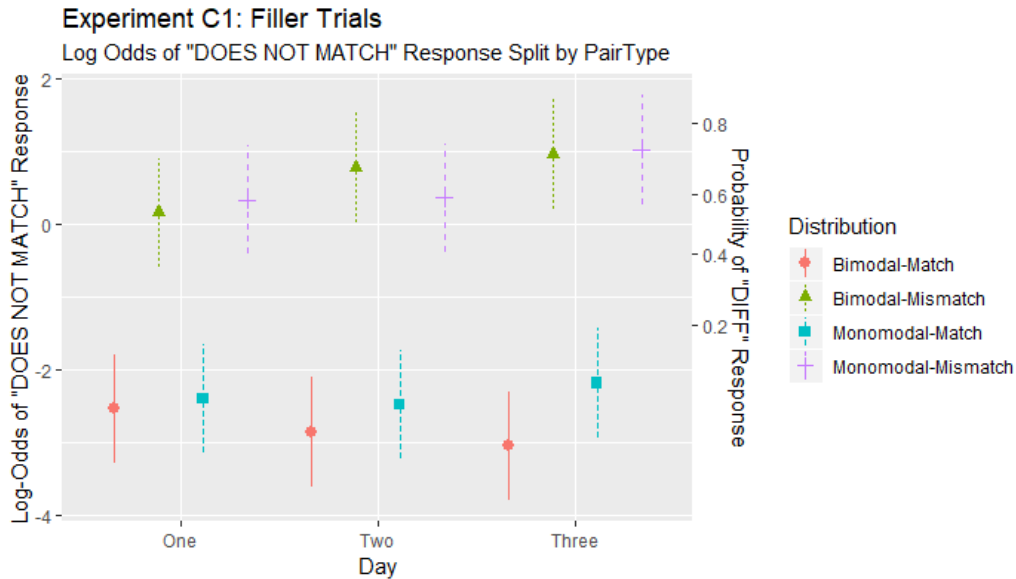


Figure 55. Results from Experiment C1 split by PairType, filler trials.

Follow-up contrasts within the context of the overall model were performed to test two specific hypotheses, on each of the three days: 1) that there is a main effect of Distribution (interpreted as **bias**, referred to here as the “Bias Test”), and 2) that there is a significant interaction between Distribution and PairType (interpreted as **sensitivity**, referred to here as the “Sensitivity Test”). Results of these follow-up contrasts are shown in Table 26. No main effect of Distribution is found on Day 1 for critical ( $p = 0.320$ ) or filler ( $p = 0.738$ ) trials; no main effect of Distribution is found on Day 2 for critical ( $p = 0.227$ ) or filler ( $p = 0.975$ ) trials; and no main effect of Distribution is found on Day 3 for critical ( $p = 0.094$ ) or filler ( $p = 0.355$ ) trials. A significant interaction is found for critical trials on Day 3, although in the opposite direction of that predicted (that is, the monomodal group has a significantly *higher* sensitivity compared to the bimodal group). Figure 56 shows results corresponding to the Bias Test, and Figure 57 shows results corresponding to the Sensitivity Test.

Day	Coefficient	SE	Wald Z	<i>p</i>
<b>BIAS: CRITICAL TRIALS</b>				
Day 1	-0.469	0.472	-0.994	0.320
Day 2	-0.743	0.615	-1.208	0.227
Day 3	-1.553	0.929	-1.673	0.094
<b>BIAS: FILLER TRIALS</b>				
Day 1	-0.162	0.485	-0.335	0.738
Day 2	0.014	0.458	0.031	0.975
Day 3	-0.462	0.500	-0.925	0.355
<b>SENSITIVITY: CRITICAL TRIALS</b>				
Day 1	-0.414	0.616	-0.672	0.501
Day 2	0.674	0.822	0.820	0.412
Day 3	-2.643	1.284	-2.058	0.040 *
<b>SENSITIVITY: FILLER TRIALS</b>				
Day 1	-0.037	0.845	-0.044	0.965
Day 2	0.796	0.885	0.900	0.368
Day 3	0.814	0.910	0.894	0.371

Table 26. Summary of follow-up contrasts testing specific hypotheses. The Bias Test tests for whether the log-odds of a Bimodal *n* response is significantly greater than the log-odds of a Monomodal *n* response. The Sensitivity Test tests for whether (log-odds of BimodalMismatched – log-odds of BimodalMatched) is significantly greater than (log-odds of MonomodalMismatched – log-odds of MonomodalMatched). Hypothesis tests were done for each of the three days.

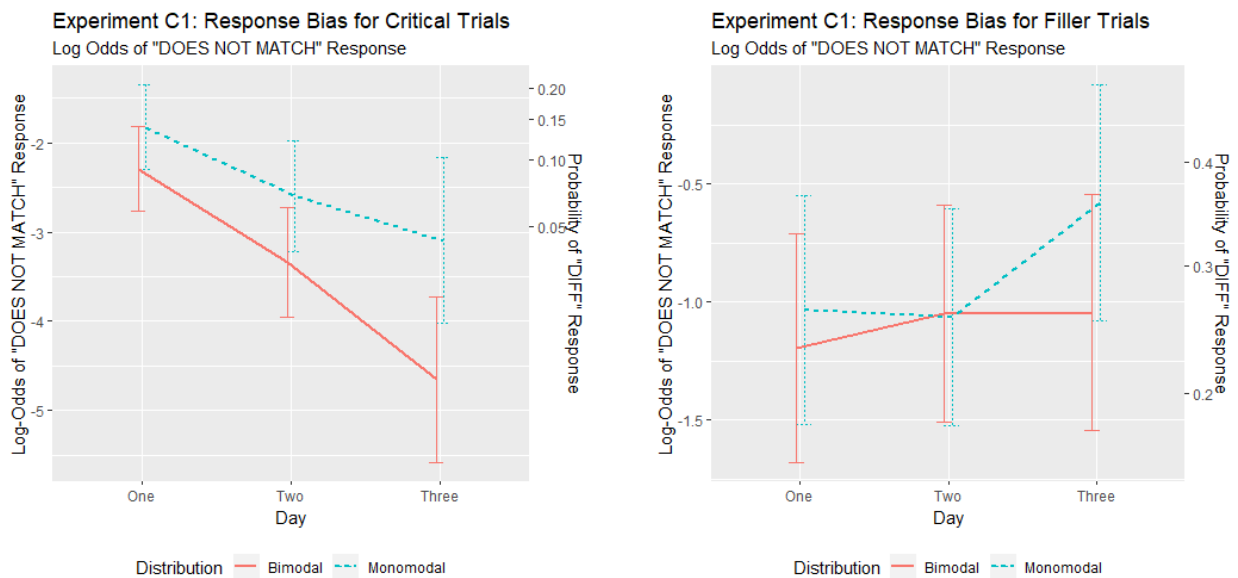


Figure 56. Bias results for Experiment C1. Figures show the total log-odds that participants responded that sound-picture pairs did not match, regardless of whether they were MatchedPairs or Mismatched-Pairs, for critical trials (left) and filler trials (right).



Figure 57. Sensitivity results for Experiment C1. Figures show the difference in log-odds for participant responses that sound-picture pairs did not match between MismatchedPairs and MatchedPairs, for critical trials (left) and filler trials (right). Higher log-odds indicate higher sensitivities (i.e. a greater distance in responses between MisMatchedPairs and MatchedPairs).

#### 4.5.2 Sleep Tests

This study also tests for any effect of sleep amount on responses, fitting a regression to the formula in (3). Since participants only received sleep-related questions when logging back into the experiment on Days 2 and 3, only responses from Days 2 and 3 are analyzed in this section. The reference cell is *bimodal, matchedPair, more*. Results of the regression are shown in Table 27.

<b>Predictor</b>	<b>Coefficient</b>	<b>SE</b>	<b>Wald Z</b>	<b>p</b>
<b>CRITICAL TRIALS</b>				
(Intercept)	-3.835	0.694	-5.530	<0.001 ***
Distribution= <i>monomodal</i>	0.884	0.814	1.086	0.277
PairType= <i>misMatchedPair</i>	-1.026	1.052	-0.975	0.330
SleepAmount= <i>less</i>	0.137	0.754	0.181	0.856
Interaction= <i>monomodal &amp; misMatchedPair</i>	0.84237	1.081	0.779	0.436
Interaction= <i>monomodal &amp; less</i>	-0.042	0.954	-0.044	0.965
Interaction= <i>misMatchedPair &amp; less</i>	-0.241	1.153	-0.209	0.834
Interaction= <i>monomodal &amp; misMatchedPair &amp; less</i>	-0.510	1.427	-0.357	0.721
<b>FILLER TRIALS</b>				
(Intercept)	-3.111	0.629	-4.950	<0.001 ***
Distribution= <i>monomodal</i>	0.902	0.688	1.312	0.190
PairType= <i>misMatchedPair</i>	4.400	0.761	5.784	<0.001 ***
SleepAmount= <i>less</i>	0.190	0.739	0.257	0.797
Interaction= <i>monomodal &amp; misMatchedPair</i>	-1.161	0.989	-1.173	0.241
Interaction= <i>monomodal &amp; less</i>	-0.552	0.951	-0.581	0.561
Interaction= <i>misMatchedPair &amp; less</i>	-0.924	0.947	-0.976	0.329
Interaction= <i>monomodal &amp; misMatchedPair &amp; less</i>	0.653	1.280	0.510	0.610

Table 27. Results of GLMM for Experiment C1, testing for an effect of sleep.

No significant interaction is found between Condition, PairType, and SleepAmount, for critical trials ( $p = 0.721$ ), or for filler trials ( $p = 0.610$ ).

Follow-up contrasts within the context of the overall model were performed to test two specific hypotheses: 1) that there is a significant interaction between Distribution and Sleep, and 2) that there is a significant main effect of Sleep. Results of the follow-up contrast testing for a significant interaction between Distribution and Sleep are shown in Table 28. Results of the follow-up contrast testing for a significant main effect of Sleep are shown in Table 29.

<b>Trial Type</b>	<b>Coefficient</b>	<b>SE</b>	<b>Wald Z</b>	<b>p</b>
Critical Trials	0.297	0.843	0.352	0.725
Filler Trials	0.226	0.582	0.388	0.698

Table 28. Summary of follow-up contrasts testing for an interaction between Distribution and Sleep, Experiment C1.

<b>Trial Type</b>	<b>Coefficient</b>	<b>SE</b>	<b>Wald Z</b>	<b>p</b>
Critical Trials	0.132	0.420	0.315	0.752
Filler Trials	0.385	0.289	1.331	0.183

Table 29. Summary of follow-up contrasts testing for a main effect of Sleep, Experiment C1.

The amount of sleep reported by the participant for the previous night does not appear to have an effect on participant response. We do not find any significant interaction between SleepAmount and Condition,

for either critical trials ( $p = 0.725$ ) or for filler trials ( $p = 0.698$ ). We also do not find any main effect of SleepAmount on Response, for either critical trials ( $p = 0.752$ ) or for filler trials ( $p = 0.183$ ).

#### 4.6. DISCUSSION: EXPERIMENT C1

Although this experiment did not directly test for distributional learning, it is assumed in this chapter that the training phase had a similar effect as that found in the “A” Experiments, since the procedure up through the training phase was identical to that of Experiment A1. If one assumes that distributional learning took place for the participants in this experiment and caused a difference in bias between the bimodal and monomodal groups (as was the case in Experiment A1), then no evidence is found that participants extended this difference in bias to a following word learning task in any of the three days tested (Figure 56). Perhaps unsurprisingly, no evidence of a difference in sensitivity between the bimodal and monomodal groups is found on Days 1 or 2. This is not surprising to find on Day 1 given that Experiment A1 concluded that participants being trained on these particular stimuli only change their biases and not their sensitivities after a single session of training. However, on Day 3, a significant difference in sensitivity between the bimodal and monomodal groups is found, in the opposite direction of that predicted by distributional learning. That is, the monomodal group has a greater sensitivity than the bimodal group in this word learning task. This study is unable to explain this difference in sensitivity.

To summarize, Experiment C1 uses stimuli that resulted in greater bias towards a “different” response in the bimodal group compared to the monomodal group after distributional training. If these changes in biases had carried over into a word-learning task, we would expect participants in the bimodal group to respond that a given sound-meaning pair does *not* match more than participants in the monomodal group. That is, they would be more hesitant to say that a sound-meaning pair which began with a “g”-like sound matched if they believed that there were two “g”-like sounds in the language, since they would be entertaining the possibility that it was the wrong “g”. Even over the course of three days, no evidence is found that any change in bias gained from distributional learning extends to a word learning task.

This study also does not find that the bimodal group exhibits greater sensitivity to critical stimuli compared to the monomodal group (although this study unexpectedly finds that the monomodal group has a greater sensitivity than the bimodal group on Day 3).

Experiment C1 tested for whether a change in response bias extends to a following word learning task, and found that it does not. For Experiment C1, this study assumed that learners trained in the bimodal group had greater bias towards a “different” response compared to learners in the monomodal group, because Experiment C1 was identical in procedure and stimuli to Experiment A1, which *did* find evidence for a change in bias. Experiment C2 will utilize a different set of critical stimuli (those used in Experiment A3). Experiment C2 will then test whether changes in sensitivities as a result of distributional learning can be extended to a word learning task.

## 5. Experiment C2

Experiment C1 finds no evidence that learners extended any sort of learned difference in their response biases to a word-learning task, even over the course of three days of training. Critical stimuli used in Experiment C1 have only been shown in this dissertation to result in bimodally-trained participants having a greater bias towards responding that a pair of critical syllables are “different” compared to monomodally-trained participants. Therefore, to test whether sensitivity shows a similar gap, Experiment C2 makes use of critical stimuli which have been shown in Chapter 3 to result in distributional learning in the form of changes in sensitivity (i.e. greater sensitivity in bimodally-trained learners than in monomodally-trained ones). Specifically, Experiment C2 tests Research Questions 2 and 3, by testing for evidence of a gap in sensitivity between a discrimination task and a word-learning task, and, if there is, test whether helps overcome this gap.

### 5.1. METHODOLOGY

The procedure of Experiment C2 follows that of Experiment C1. The main difference is that critical stimuli consist of [eɑ] – [ʒɑ], rather than [gɑ gæ gɹ] – [gɑ gæ gɹ].

### 5.1.1 Stimuli

Critical stimuli were based off those used in Experiment A3. No-coda stimuli were identical to those used in Experiment B, and ranged from [ɛa] – [ʂa] (see Figure 58). As with Experiment C1, coda critical stimuli were created by splicing shortened no-coda critical stimuli with intensity-adjusted codas. To obtain shortened no-coda critical stimuli, approximately 20 ms were removed from the ends of all no-coda critical stimuli (where splice points were made at positive-going zero crossings). Intensity-adjusted codas originated from the following recorded syllables: [ʃamp], [ʃasp], [ʃaŋ], [ʃavz], [ʃad], and [ʃab]. Codas from each of these syllables were removed, and their intensities were modified in Praat. The amount of intensity-modification was chosen through trial and error so that splicing no-coda critical stimuli with codas sounded natural. Note that while the critical stimuli in Experiment C1 were made up of three different nucleus types ([a æ ɪ]), the stimuli in this experiment all had the same [a] nucleus. No-coda filler stimuli were identical to those used in Experiment C1.

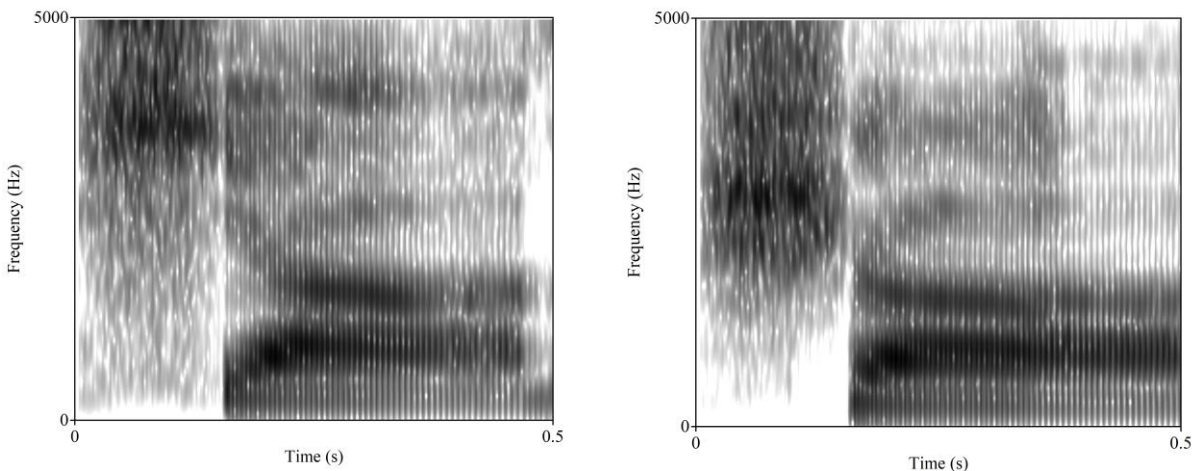


Figure 58. First 500 ms of critical syllables S<sub>1a</sub> (left) and S<sub>8a</sub> (right).

### 5.1.2 Procedure

Participants were again recruited through Mechanical Turk. This experiment took place over the course of three days. As the case with Experiment C1, Experiment C2 consisted of 8 phases, only some of which

were presented on each day of the experiment. A summary of the phases presented on each day is shown in Figure 59.

<b>Day 1</b>	<b>Day 2</b>	<b>Day 3</b>
Login	Login	Login
---	Sleep questionnaire	Sleep questionnaire
Sound check	Sound check	Sound check
English practice word test	English practice word test	English practice word test
Training	Training	Training
Word learning	Word learning	Word learning
Word test	Word test	Word test
---	---	Questionnaire

Figure 59. Summary of procedure on each day. The double-boxed portion indicates a procedure identical to that found in Experiment A1 before the test phase.

Each day began with a **login**, which prompted users to enter their Mechanical Turk ID. On Days 2 and 3, this was followed by a **sleep questionnaire**, which asked participants how many hours they had slept the previous night, and how they would rate the quality of their sleep. Following this was a **sound check**, identical to that in Experiment C1.

During each block of **training**, participants heard three repetitions of 16 tokens drawn from the critical S<sub>1</sub>a-S<sub>8</sub>a continuum, for a total of 48 critical no-coda syllables per block. Additionally, they heard 4 different tokens of 4 filler syllables ([t<sup>h</sup>ɑ], [tɑ], [fɑ], [hɑ]). Each of these 16 filler tokens were repeated 3 times during each train block, for a total of 48 fillers per block. Training on each day consisted of 4 blocks, resulting in a total of 192 critical syllables and 192 filler syllables per day.

Following training, participants were directed to a **word-learning phase**. The word-learning phase was identical to that presented to participants in Experiment C1, but used different critical stimuli. Sound-meaning pairs in each subcondition are shown in Figure 60.



Subcondition A	Subcondition B	Meaning
S <sub>1-4</sub> amp	S <sub>5-8</sub> amp	apple
S <sub>5-8</sub> asp	S <sub>1-4</sub> asp	fork
S <sub>1-4</sub> aŋ	S <sub>5-8</sub> aŋ	chair
S <sub>5-8</sub> avz	S <sub>1-4</sub> avz	boot
S <sub>1-4</sub> ad	S <sub>5-8</sub> ad	elephant
S <sub>5-8</sub> ab	S <sub>1-4</sub> ab	lamp
fæs <sub>1</sub> , fæs <sub>2</sub>	fæs <sub>1</sub> , fæs <sub>2</sub>	motorcycle
tɛb <sub>1</sub> , tɛb <sub>2</sub>	tejb <sub>1</sub> , tejb <sub>2</sub>	onion
mæfs <sub>1</sub> , mæfs <sub>2</sub>	næfs <sub>1</sub> , næfs <sub>2</sub>	horse
sɛm <sub>1</sub> , sɛm <sub>2</sub>	zɛm <sub>1</sub> , zɛm <sub>2</sub>	trumpet

Figure 60. Sound-meaning pairs presented during the word learning phase.

As was the case for Experiment C1, each trial consisted of an auditory stimulus and a visual presented on the screen for 2000 ms, before the next trial began. Participants were instructed to simply try to memorize the words presented in this phase, and were reminded not to write anything down.

Following the word learning phase, participants were directed to a test phase. Directions and procedure were identical to those given in Experiment C1. Again, in the word test phase, participants were exposed to sound-picture pairs that were either MatchedPairs or MismatchedPairs.

Meaning	Matched (for Subcond. A)	Mismatched (for Subcond. A)
	Mismatched (for Subcond. B)	Matched (for Subcond. B)
apple	S <sub>1</sub> amp	S <sub>8</sub> amp
fork	S <sub>8</sub> asp	S <sub>4</sub> asp
chair	S <sub>1</sub> aŋ	S <sub>8</sub> aŋ
boot	S <sub>8</sub> avz	S <sub>4</sub> avz
elephant	S <sub>1</sub> ad	S <sub>8</sub> ad
lamp	S <sub>8</sub> ab	S <sub>1</sub> ab
motorcycle	fæs	fæs
onion	tɛb	tejb
horse	mæfs	næfs
trumpet	sɛm	zɛm

Figure 61. Summary of MatchedPairs and MismatchedPairs.

A brief summary of each phase presented in this experiment is shown in Figure 62.

<b>Phase</b>	<b>Abbreviated directions</b>	<b>Example</b>
Login		
Sleep questionnaire (only on Days 2 and 3)	<i>How well did you sleep? How much sleep did you get?</i>	
Sound check	<i>Press the '1' if you hear 1 beep, the '2' if you hear 2</i>	
English practice word test	<i>Press 'm' if the sound <u>m</u>atches the picture you see, 'n' if they do <u>n</u>ot match</i>	<i>disk</i> (see picture of a desk)
Training	<i>Listen carefully</i>	<i>S<sub>3a</sub></i> <i>fa</i>
Word learning	<i>Try to memorize these words</i>	<i>S<sub>1-amp</sub></i> (see picture of apple)
Test	<i>Press 'm' if the sound <u>m</u>atches the picture you see, 'n' if they do <u>n</u>ot match</i>	<i>S<sub>8amp</sub></i> (see picture of apple)
Questionnaire (only on Day 3)		

Figure 62. Summary of each phase in Experiment C2.

A total of 104 participants participated on the first day. 68 of these participants took all three days of the experiment. Because the questionnaire took place on the third day, only participants who had participated all three days were considered for analysis. Participants were excluded from analysis if they reported in their questionnaire answers that they: (1) had a history of a speech or hearing disorder (0 participants excluded for this reason); (2) were not native speakers of English (0 participants excluded for this reason); (3) wrote down words (0 participants excluded for this reason); or (4) reported having some background in a language with two or more voiceless post-alveolar fricatives as phonemes (10 participants excluded). A copy of the questionnaire is included in the Appendix.

Participants were also excluded if they: (1) received a combined score across all three days of less than 15/18 on the sound check (12 participants excluded for this); or (2) a combined score across all three days of less than 15/24 on the English practice test (0 excluded for this). This left a total of 47 participants (27 bimodal, 20 monomodal), with some participants excluded for multiple reasons.

## 5.2. ANALYSIS AND RESULTS

### 5.2.1 Bias and Sensitivity Tests

Experiment C2 followed the analysis used in Experiment C1 (see Section 4.4 for details). The results of the GLMM are shown in Table 30. Figure 63 (critical trials) and Figure 64 (filler trials) display results of Experiment C2 split by PairType.

Predictor	Coefficient	SE	Wald Z	p
<b>CRITICAL TRIALS</b>				
(Intercept)	-1.987	0.303	-6.560	<0.001 ***
Distribution= <i>monomodal</i>	0.180	0.386	0.467	0.641
PairType= <i>misMatchedPair</i>	0.596	0.365	1.633	0.103
Day= <i>two</i>	0.050	0.386	0.129	0.897
Day= <i>three</i>	-0.541	0.395	-1.369	0.171
Interaction= <i>monomodal &amp; misMatchedPair</i>	0.017	0.532	0.032	0.975
Interaction= <i>monomodal &amp; two</i>	-0.717	0.573	-1.252	0.211
Interaction= <i>monomodal &amp; three</i>	-0.385	0.597	-0.644	0.520
Interaction= <i>misMatchedPair &amp; two</i>	-0.287	0.450	-0.638	0.524
Interaction= <i>misMatchedPair &amp; three</i>	0.417	0.467	0.893	0.372
Interaction= <i>monomodal &amp; misMatchedPair &amp; two</i>	0.080	0.700	0.114	0.909
Interaction= <i>monomodal &amp; misMatchedPair &amp; three</i>	0.106	0.716	0.149	0.882
<b>FILLER TRIALS</b>				
(Intercept)	-1.722	0.387	-4.450	<0.001 ***
Distribution= <i>monomodal</i>	-0.327	0.475	-0.687	0.492
PairType= <i>misMatchedPair</i>	1.435	0.434	3.305	0.001
Day= <i>two</i>	-1.126	0.459	-2.451	0.014
Day= <i>three</i>	-1.191	0.482	-2.470	0.014
Interaction= <i>monomodal &amp; misMatchedPair</i>	0.237	0.656	0.362	0.718
Interaction= <i>monomodal &amp; two</i>	0.858	0.673	1.276	0.202
Interaction= <i>monomodal &amp; three</i>	-0.225	0.831	-0.270	0.787
Interaction= <i>misMatchedPair &amp; two</i>	1.576	0.548	2.876	0.004
Interaction= <i>misMatchedPair &amp; three</i>	2.459	0.572	4.301	<0.001 ***
Interaction= <i>monomodal &amp; misMatchedPair &amp; two</i>	-0.836	0.813	-1.028	0.304
Interaction= <i>monomodal &amp; misMatchedPair &amp; three</i>	0.122	0.955	0.128	0.898

Table 30. Results of GLMM for Experiment C2.

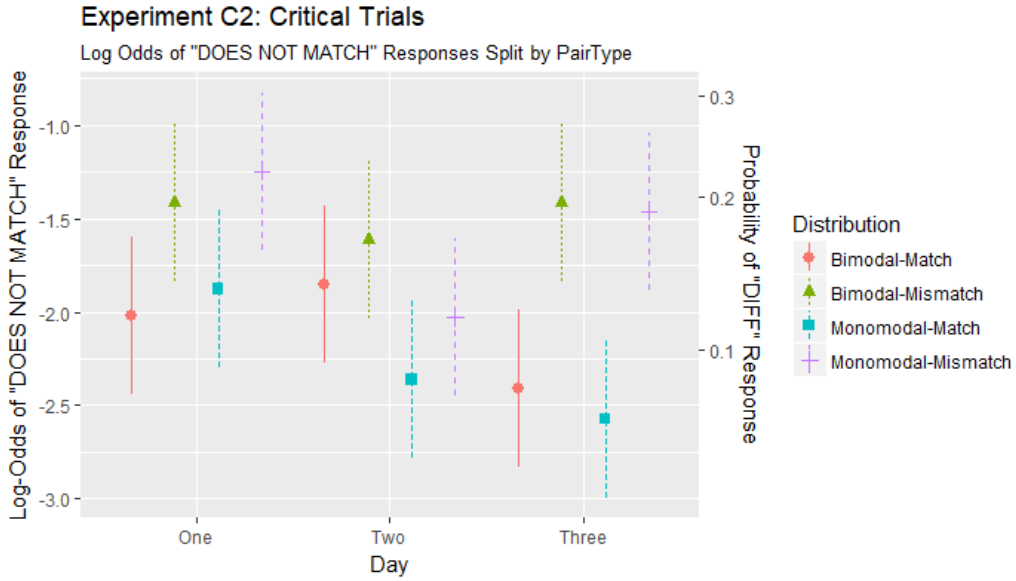


Figure 63. Results from Experiment C2 split by PairType, critical trials.

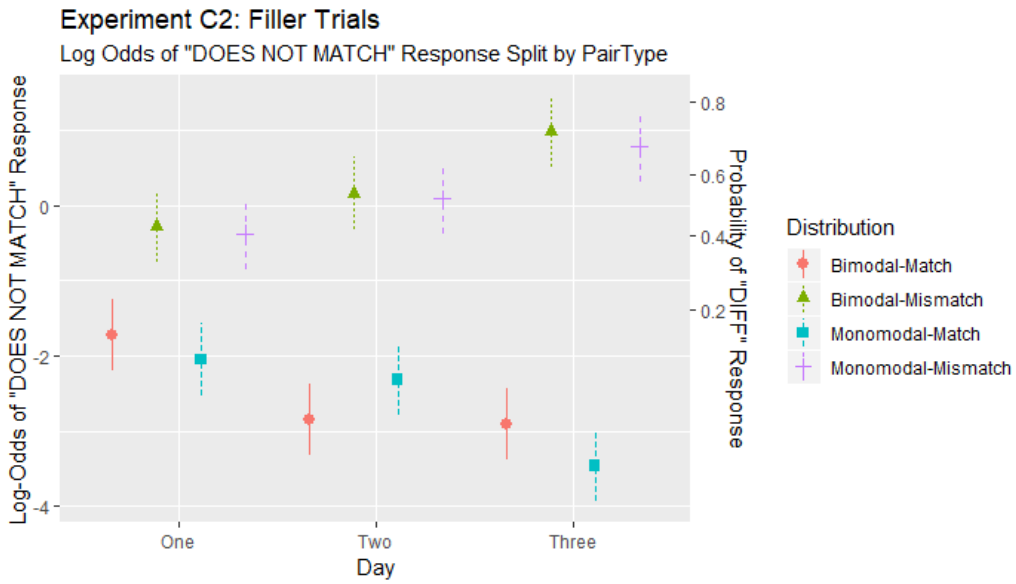


Figure 64. Results from Experiment C2 split by PairType, filler trials.

Follow-up contrasts were done in the context of the overall model to test two specific hypotheses: 1) that there was a main effect of Distribution (Bias Test), and 2) that there is a significant interaction between Distribution and PairType (Sensitivity Test). These hypotheses were tested on each of the three days. Results are shown in Table 31.

<b>Day</b>	<b>Coefficient</b>	<b>SE</b>	<b>Wald Z</b>	<b><i>p</i></b>
<b>BIAS: CRITICAL TRIALS</b>				
Day 1	-0.188	0.317	-0.595	0.552
Day 2	0.489	0.403	1.214	0.225
Day 3	0.143	0.447	0.320	0.749
<b>BIAS: FILLER TRIALS</b>				
Day 1	0.208	0.312	0.667	0.505
Day 2	-0.232	0.366	-0.634	0.526
Day 3	0.371	0.444	0.837	0.402
<b>SENSITIVITY: CRITICAL TRIALS</b>				
Day 1	-0.017	0.532	-0.032	0.975
Day 2	-0.097	0.589	-0.164	0.869
Day 3	-0.123	0.610	-0.202	0.840
<b>SENSITIVITY: FILLER TRIALS</b>				
Day 1	-0.237	0.656	-0.362	0.718
Day 2	0.599	0.728	0.823	0.411
Day 3	-0.359	0.882	-0.408	0.683

Table 31. Summary of follow-up contrasts testing specific hypotheses. The Bias Test tests for whether the log-odds of a Bimodal *n* response is significantly greater than the log-odds of a Monomodal *n* response. The Sensitivity Test tests for whether (log-odds of BimodalMismatched – log-odds of BimodalMatched) is significantly greater than (log-odds of MonomodalMismatched – log-odds of MonomodalMatched). Hypothesis tests were done for each of the three days.

No significant main effect of Distribution is found on any of the three days. No significant interaction between Distribution and PairType is found on any of the three days. Figure 65 shows results corresponding to the Bias Test. Figure 66 shows results corresponding to the Sensitivity Test.

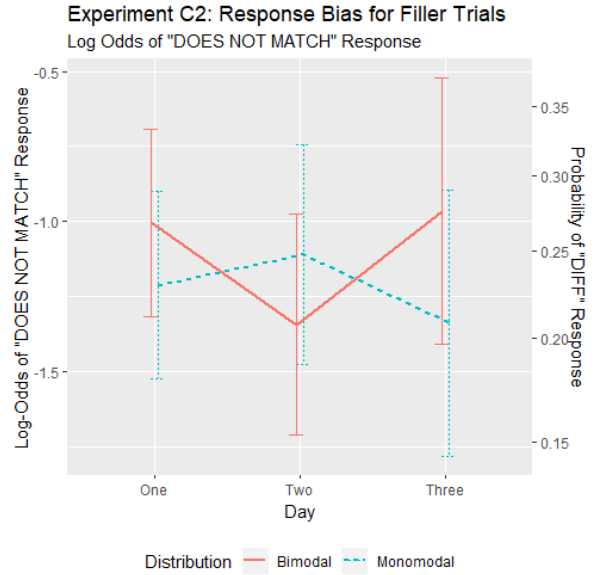
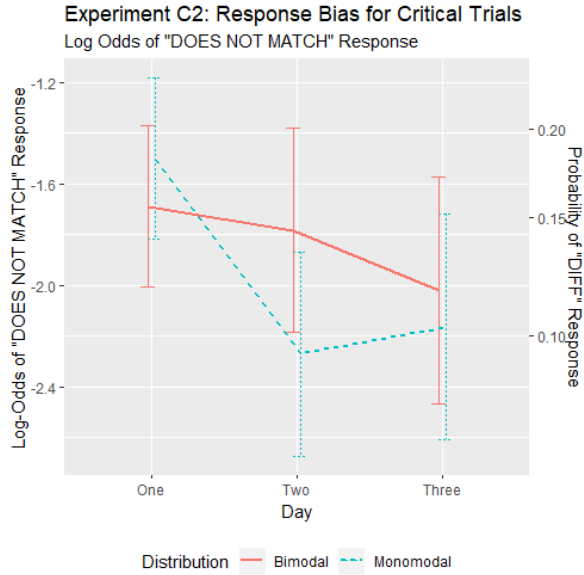


Figure 65. Bias results for Experiment C1. Figures show the total log-odds that participants responded that sound-picture pairs did not match, regardless of whether they were MatchedPairs or Mismatched-Pairs, for critical trials (left) and filler trials (right).

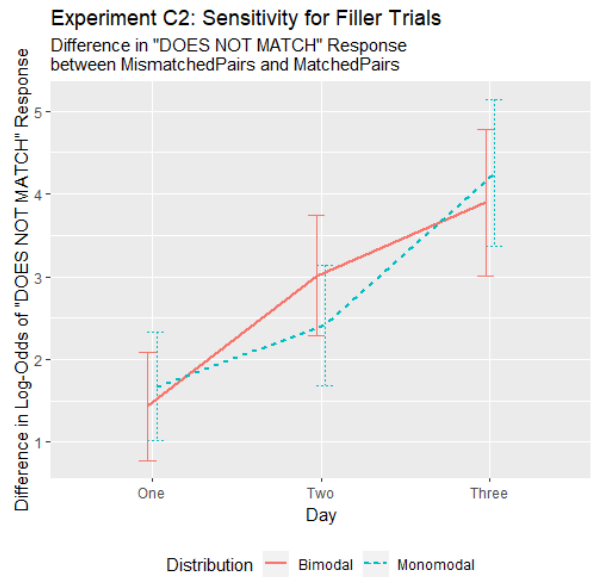
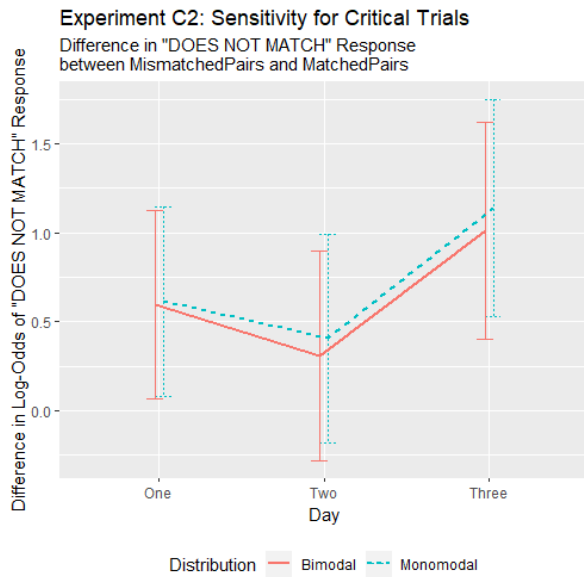


Figure 66. Sensitivity results for Experiment C1. Figures show the difference in log-odds for participant responses that sound-picture pairs did not match between MismatchedPairs and MatchedPairs, for critical trials (left) and filler trials (right). Higher log-odds indicate higher sensitivities (i.e. a greater distance in responses between MismatchedPairs and MatchedPairs).

### 5.2.2 Sleep Tests

As with Experiment C1, this study also tests for any effect of sleep amount on responses, fitting a regression to the formula in (3). Since participants only received sleep-related questions when logging back into the experiment on Days 2 and 3, only responses from Days 2 and 3 are analyzed in this section. The reference cell is *bimodal, matchedPair, more*. Results of the regression are shown in Table 32.

Predictor	Coefficient	SE	Wald Z	p
<b>CRITICAL TRIALS</b>				
(Intercept)	-2.345	0.396	-5.921	<0.001 ***
Distribution= <i>monomodal</i>	-0.505	0.593	-0.852	0.394
PairType= <i>misMatchedPair</i>	0.592	0.404	1.468	0.142
SleepAmount= <i>less</i>	-0.086	0.447	-0.193	0.847
Interaction= <i>monomodal &amp; misMatchedPair</i>	0.183	0.632	0.289	0.773
Interaction= <i>monomodal &amp; less</i>	0.036	0.743	0.048	0.962
Interaction= <i>misMatchedPair &amp; less</i>	0.447	0.529	0.845	0.398
Interaction= <i>monomodal &amp; misMatchedPair &amp; less</i>	0.032	0.847	0.038	0.970
<b>FILLER TRIALS</b>				
(Intercept)	-3.149	0.559	-5.628	<0.001 ***
Distribution= <i>monomodal</i>	0.374	0.675	0.554	0.579
PairType= <i>misMatchedPair</i>	3.673	0.646	5.682	<0.001 ***
SleepAmount= <i>less</i>	-0.006	0.626	-0.010	0.992
Interaction= <i>monomodal &amp; misMatchedPair</i>	-0.529	0.893	-0.592	0.554
Interaction= <i>monomodal &amp; less</i>	-0.421	0.941	-0.448	0.654
Interaction= <i>misMatchedPair &amp; less</i>	0.138	0.772	0.179	0.858
Interaction= <i>monomodal &amp; misMatchedPair &amp; less</i>	0.476	1.173	0.406	0.685

Table 32. Results of GLMM for Experiment C2, testing for an effect of sleep.

No significant interactions between Condition, PairType, and SleepAmount are found for critical trials ( $p = 0.970$ ), or for filler trials ( $p = 0.685$ ).

Follow-up contrasts within the context of the overall model were performed to test two specific hypotheses: 1) that there is a significant interaction between Distribution and Sleep, and 2) that there is a significant main effect of Sleep. Results of the follow-up contrast testing for a significant interaction between Distribution and Sleep are shown in Table 33. Results of the follow-up contrast testing for a significant main effect of Sleep are shown in Table 34.

Trial Type	Coefficient	SE	Wald Z	p
Critical Trials	-0.052	0.528	-0.098	0.922
Filler Trials	0.183	0.562	0.326	0.744

Table 33. Summary of follow-up contrasts testing for an interaction between Distribution and Sleep, Experiment C1.

<b>Trial Type</b>	<b>Coefficient</b>	<b>SE</b>	<b>Wald Z</b>	<b><i>p</i></b>
Critical Trials	-0.163	0.268	-0.609	0.542
Filler Trials	0.029	0.278	0.103	0.918

Table 34. Summary of follow-up contrasts testing for a main effect of Sleep, Experiment C2.

The amount of sleep reported by the participant for the previous night does not appear to have an effect on participant response. We do not find any significant interaction between SleepAmount and Condition, for either critical trials ( $p = 0.922$ ) or for filler trials ( $p = 0.744$ ). We also do not find any main effect of SleepAmount on Response, for either critical trials ( $p = 0.542$ ) or for filler trials ( $p = 0.918$ ).

### 5.3. DISCUSSION: EXPERIMENT C2

Experiment C2 finds no evidence that any difference in sensitivity assumed to exist between bimodally-trained participants and monomodally-trained participants (where this assumption is based on results of Experiment A3) carries over to a word learning task, even over the course of three days of training and testing.

## 6. Discussion

This chapter seeks to determine whether knowledge gained from distributional learning carries over to a word-learning task. In two three-day long experiments, this study finds no evidence that either of the two aspects of distributional learning (changes in bias or sensitivity) extend to word learning.

Although proposals have been made regarding the gap in sensitivity from discrimination to word learning tasks (Werker and Curtin, 2005; Pater et al., 2004), these proposals do not predict a gap in response bias as well. Based on results of this study, this chapter argues that these proposals should be extended so that attention levels also play a role in response bias: specifically, the greater cognitive demands of a word learning task result in less attention to phonetic detail as well as less explicit inference from the participant that there are multiple contrastive phonemes (and therefore less inference that minimal pairs exist) in the speech signal.



Additionally, this study does not find support for the hypothesis that the participants in an artificial word learning task similar to Hayes-Harb (2007) simply needed a period of sleep to integrate newly-learned phonetic category information in a word learning task, since this experiment finds no evidence that sensitivity or response bias extended to word learning even on Days 2 and 3.

## **7. Conclusion**

If participants are assumed to have reached the bias stage in distributional learning by the time of the word-learning phase on Day 1 of Experiment C1, then this study finds no evidence that these biases extend to word learning. Although it is unknown what stage of distributional learning learners in Experiment C1 have reached on Days 2 and 3, this study still does not find evidence that the response bias from Day 1 has extended to a word learning phase on these days. Similarly, if it is assumed that participants have reached the sensitivity stage in distributional learning by the time of the word-learning phase on Day 1 of Experiment C2, then this study finds no evidence that this sensitivity extends to word learning, even on Days 2 and 3. This lack of evidence suggests that there is both a gap in sensitivity, as well as in response bias, when participants are faced with the more cognitively-demanding nature of a word learning task. This is the case even after participants have had a chance to sleep, contradicting the hypothesis that a period of sleep is all that is necessary to integrate phonetic category acquisition with the application of phonetic categories to word learning.

## **Chapter 6:**

### **Discussion and Conclusion**

#### **1. Introduction**

The overall goal of this dissertation has been to provide a detailed timeline of segmental acquisition, starting from learning phonetic categories, to determining allophonic relationships between phonetic categories, to utilizing sound categories in a meaningful, linguistic way in word learning. This was done through a series of artificial language learning experiments on naïve adult listeners recruited over the web using Mechanical Turk. This chapter begins with a summary of significant findings from Experiments A-C in Section 2. This will be followed in Section 3 by an outline of the overall proposed timeline of segmental acquisition. Section 4 will address outstanding questions and will attempt to shed light on unexplained (and unexpected) significant results, as well as suggest topics for further research.

#### **2. Summary of Findings**

The main contribution of this dissertation has been to make a distinction between response bias and sensitivity in phonetic category acquisition. Table 35 provides a summary of all significant findings, either in bias or in sensitivity, in Experiments A-C.

Experiment	Critical Stimuli	Filler Stimuli
A1	Bias ( <i>Bi &gt; Mono</i> )	Sensitivity ♦ ( <i>Mono &gt; Bi</i> )
A2	Bias ( <i>Bi &gt; Mono</i> )	---
A3	Sensitivity ( <i>Bi &gt; Mono</i> )	Sensitivity ♦ ( <i>Bi &gt; Mono</i> )
A2-Tone	---	---
A3-Tone	---	---

Experiment		Critical Stimuli	Filler Stimuli
B (Phone Test)	Time 1	Bias ( <i>Bi-NonComp &gt; Mono</i> )	
	Time 2	Sensitivity ( <i>Mono &gt; Bi-Comp</i> )	Sensitivity ♦ ( <i>Bi-Comp &gt; Bi-NonComp</i> ) ♦ ( <i>Mono &gt; Bi-NonComp</i> )
	Time 3	Sensitivity ( <i>Bi-NonComp &gt; Bi-Comp</i> )	

Experiment		Old Stimuli	New Stimuli
B (Rule Test)	Time 1	Sensitivity ( <i>Bi-Comp &gt; Bi-NonComp</i> )	---
	Time 2	Sensitivity ( <i>Bi-Comp &gt; Mono</i> )	---
	Time 3	---	---

Experiment		Critical Stimuli	Filler Stimuli
C1	Day 1	---	---
	Day 2	---	---
	Day 3	Sensitivity ♦ ( <i>Mono &gt; Bi</i> )	---
C2	Day 1	---	---
	Day 2	---	---
	Day 3	---	---

Table 35. Summary of significant results in Experiments A-C. Diamonds indicate unexpected/unexplained findings.

This section will briefly summarize conclusions based on these results. Findings which were unexpected and are so far unexplained are marked with diamonds. These will be discussed further in this chapter.

Experiments A1 and A2 found a significantly greater bias towards a “different” response for bimodally-trained participants compared to monomodally-trained ones, while Experiment A3 found a

significantly greater sensitivity in bimodally-trained participants compared to monomodally-trained ones. Together, the results from Experiments A1-A3 were used to argue for a two-stage process for distributional learning of phonetic categories. First, participants undergo a Bias Stage, in which their attention is drawn to variation in the input. This has the effect of creating general rough ideas regarding the number of distinct sound categories in the input. Over time, the acoustic location of the mean and boundaries of each phonetic category become more solid. Following this, participants experience perceptual warping in a Sensitivity Stage, where acoustic stimuli deemed to belong to the same category are perceived as being more similar to one another and acoustic stimuli deemed to belong to separate categories are perceived as being more distinct from one another. This model does not require language-specific cognitive mechanisms, and instead can be explained using domain-general processing.

Experiments A2-Tone and A3-Tone followed up on Experiments A2 and A3, only differing in the inclusion of low-frequency tones interspersed with speech stimuli during the Train phase. The inclusion of these tones resulted in non-results for both Experiments A2-Tone and A3-Tone. These non-results were used to argue that attention plays a role in distributional learning, although the exact nature of that role is left for future research.

The main goal of Experiment B was to determine whether experimental evidence could be found for a two-stage or a one-stage model of allophony acquisition. Experiment B contained two tests: a Phone Test and a Rule Test. Participants were exposed to one of three lengths of training, where training was either monomodal, bimodal and not in complementary distribution (“Bimodal-NonComp”), or bimodal and in complementary distribution (“Bimodal-Comp”). At the longest exposure time (Time 3), Experiment B finds that the Bimodal-NonComp group has a significantly greater sensitivity to critical stimuli than the Bimodal-Comp group, replicating results from Noguchi (2016). At the middle exposure time (Time 2), the Monomodal group exhibited significantly higher sensitivity to critical stimuli than the Bimodal-Comp group. Although these results are not clear evidence for either the one- or the two-stage model of allophony acquisition, Chapter 4 concludes that these results are better explained with a one-

stage model rather than a two-stage one, since at no point in time does the Bimodal-Comp group exhibit greater numeric sensitivity to critical stimuli than the Monomodal group, even exhibiting significantly less sensitivity to critical stimuli than the Monomodal group at Time 2. Additionally, Experiment B finds further support for the two-stage model of phonetic category acquisition provided in Chapter 3, as evidence for greater bias towards a “different” response is found in the Bimodal-NonComp group compared to the Monomodal group at Time 1.

The Rule Test of Experiment B shows 1) a lack of generalization of the phonological rule to new syllables, and 2) a possible effect of fatigue. As only the Bimodal-Comp group was exposed to the phonological rule that [ɛ] occur after [i] and [ɤ] occur after [u], only the Bimodal-Comp group was expected to show sensitivity to this rule. The phonological rule was tested with both stimuli experienced during training (Old stimuli), as well as stimuli not heard during training (New stimuli). The Bimodal-Comp group does show the expected sensitivity to the phonological rule they had been trained on at Times 1 and 2. Specifically, the Bimodal-Comp group shows greater sensitivity to the difference between phonologically-legal and phonologically-illegal phrases than the Bimodal-NonComp group at Time 1, as well as the Monomodal group at Time 2. There is no evidence that participants generalized this rule to new syllables though, as only Old stimuli yielded this significant difference. No significant differences were found in New stimuli. A possible effect of participant fatigue can be seen at Time 3, as even with the most training, the Bimodal-Comp group does not show significantly greater sensitivity to phonologically-legal and illegal phrases compared with either of the other groups.

Finally, the “C” Experiments tested whether changes in bias (Experiment C1) or sensitivity (Experiment C2) caused by distributional learning could be extended to a word learning task. Neither experiment shows any evidence of this being the case, although Experiment C1 had the unusual finding that participants in the Monomodal condition exhibited greater sensitivity to stimuli than those in the Bimodal condition on Day 3. Although further research is necessary to determine the cause of this, a

breakdown of results of Experiment C1 seem to indicate that this unusual result seems to come solely from the behavior of Bimodal participants on Day 3 responding to Mismatched trials (see Figure 67).

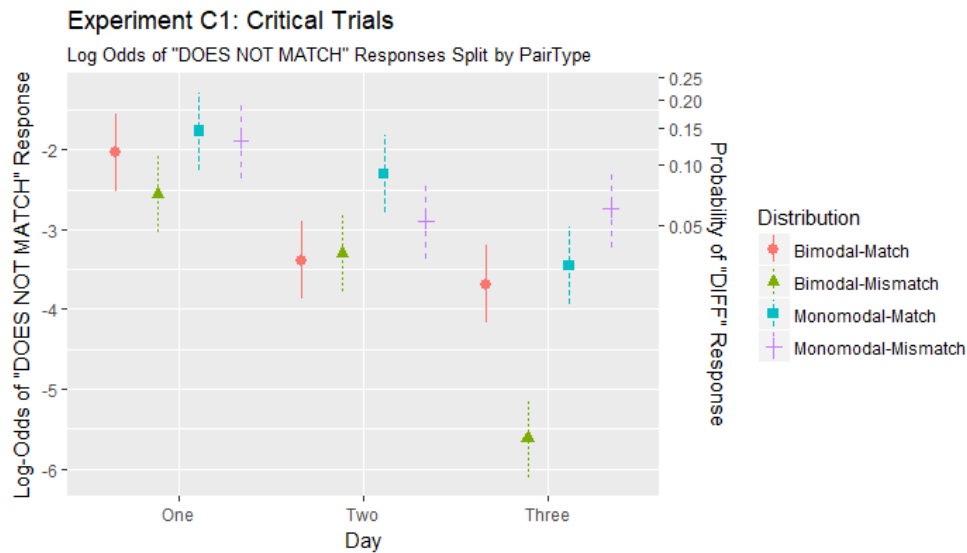


Figure 67. Breakdown of results from Experiment C1, critical trials. Note the location of the Bimodal-Mismatch point on Day 3.

As seen in Figure 67, the Bimodal group on Day 3 is very unlikely to (correctly) respond that Mismatch trials (i.e. pictures which did match the wordform they had originally been trained on) did not in fact match the correct pronunciation. Further research is necessary to explain this result, but it is thought that this is a chance occurrence isolated to the Bimodal group.

The following section takes a step back from the data and provides a summary of the overall proposal presented in this dissertation.

### 3. Synopsis of the Proposal

The overall contribution of this dissertation is to provide a timeline of early segmental acquisition. A schematic of this proposal is shown in Figure 68.

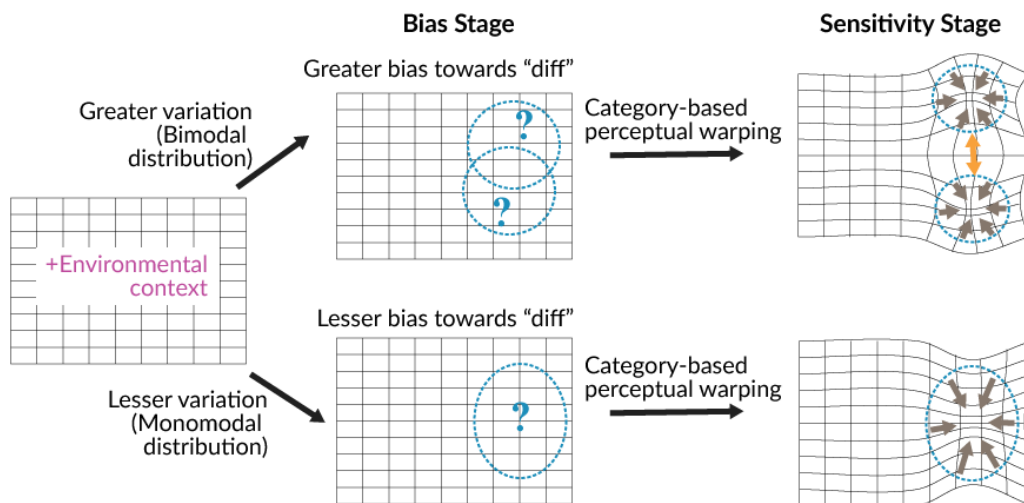


Figure 68. Schematic of overall proposal.

The grid here represents the learner’s perceptual space. The initial stage of acquisition is represented by the leftmost grid in Figure 68. This dissertation proposes that language learners who notice greater variation will have a greater **bias** toward believing that there are two different sound categories in the speech stream, compared to learners who do not notice as much variation. Specifically, Chapter 3 argues that learners trained on phones ranging along an 8-point continuum whose frequencies form a bimodal distribution will be more likely to hear continuum points 2 and 7 side by side more often than learners who hear these same continuum points in a frequency that forms a monomodal distribution. Monomodally-trained learners on the other hand will be more likely to hear continuum points 4 and 5 side by side. Since the points 2 and 7 are more acoustically distinct from one another, bimodally-trained learners are more likely to notice variation between the two tokens compared to monomodally-trained learners. This raised awareness of variation in the speech stream in turn makes learners either more likely to think there are more phones in the speech stream, or more likely to think that variation must be linguistically significant and not just a chance occurrence. Learners trained on the bimodal distribution form a rough notion that there are two phonetic categories, whereas learners trained on the monomodal distribution form a rough notion that there is only one. Crucially, at the Bias Stage, the identity of each phonetic category is only a rough notion to the learner, as indicated by the dotted lines. It is only later during the Sensitivity Stage

that the learner forms a more solid concept of the distinction between each phonetic category. At this stage, learners experience category-based perceptual warping. Specifically, sounds deemed by the learner to belong to the same category are perceived as being more similar to one another through **within-category compression**, and sounds deemed by the learner to belong to different categories are perceived as being more different from one another through **across-category expansion**. The concepts of within-category compression and across-category expansion are borrowed from the psychology literature (e.g. Livingston et al. 1998; Goldstone and Hendrickson, 2010).

This proposal contrasts with an experience-based perceptual warping proposal illustrated in Figure 69. Here, the experience of perceiving a token warps the learner’s perceptual space. Experience-based perceptual warping proposals have been made by researchers such as Guenther and Gjaja (1996) and Borsma et al. (2003).

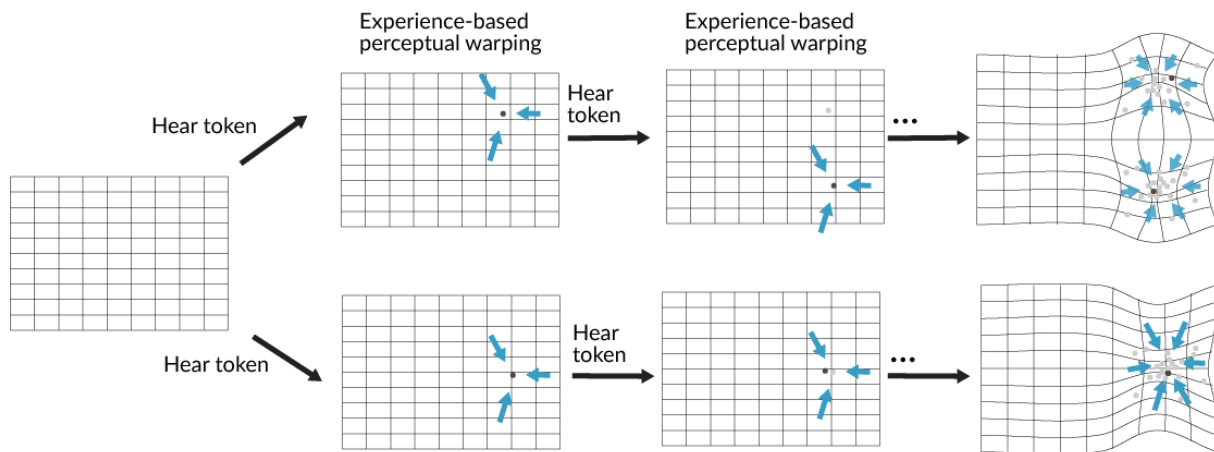


Figure 69. Experience-based perceptual warping account of phonetic category acquisition.

This overall proposal is discussed in terms of its interaction with three main topics: attention, environmental context, and lexical acquisition. Experiments A2-Tone and A3-Tone provide evidence that attention plays some role in the *magnitude* of the arrows shown in leading up to the Bias Stage, and possibly also the magnitude of the arrows leading up to the Sensitivity Stage. Experiment B provides some evidence that environmental context is accounted for early on, rather than at a later stage of acquisition



occurring after the acquisition of phonetic categories. And last, the “C” Experiments find no evidence that changes in bias or sensitivity extend to a word learning task, even after three consecutive days of training.

#### **4. Discussion and Further Research**

Results of this dissertation raise a number of questions for further research. These will be discussed in this section.

##### **4.1. BEHAVIOR OF FILLER STIMULI**

One of the main problems with the experiments reported here is that a number of them yielded significant results for the *filler* stimuli. Ideally, these stimuli should be able to act as controls, in order to determine whether there is something unexpected in the actual design of the experiment. One could argue that these significant results with the filler items indicate that there is some underlying problem with the design of the experiment. This may be the case, but I believe the behavior with the fillers can be explained by the nature of the “same” stimuli I created for these experiments. All experiments which exhibited unexpected significant results for filler trials were experiments for which I had created the stimuli, rather than use those created by a previous researcher. As explained in the discussion section of Chapter 3, my “same” fillers were created with the thought that these *same-different* experiments should be differentiated from discrimination tasks, and instead should require participants to make some sort of explicit decision regarding a pair of stimuli’s “sameness” (in keeping with discussion provided in Maye and Gerken, 2000). Therefore, I purposely recorded repetitions of “same” syllables so that they could be distinguished by ear as being different tokens. This differed from the stimuli provided by Maye and Gerken. To my ears, repetitions of “same” filler tokens used by Maye and Gerken were so similar as to be perceptually identical. Because of this, it would be unlikely that any participant would respond that a filler Same Pair was actually “different.” I believe the filler stimuli used by Maye and Gerken exhibited a floor effect for this reason. As for why my fillers exhibited the unexpected behavior of being seemingly affected by condition despite being presented identically in bimodal and monomodal conditions, it may be the case that training had an effect on all stimuli. It may be the case that participants were in essence calibrating what counted

as a simple variation in pronunciation and what counted as a linguistically-meaningful replacement of phonemes during training. Training had the effect of shrinking or stretching what counted as “meaningless” variation, and participants used this distance on all stimuli, not just those responsible for the shrinking or stretching. If a pair of stimuli, critical or not, fit the learner’s current hypothesis of being “meaningless” variation, they would respond that the pair of syllables were the “same.” If not, they would respond “different.” This basic concept is captured by the “yardstick model” presented in DeCarlo (2013), who presents several models for translating Signal Detection Theory concepts into a *same-different* experiment. I believe it would be interesting to conduct future distributional learning work within the framework of this model.

#### 4.2. LACK OF SIGNIFICANT FINDINGS IN “C” EXPERIMENTS

The “C” Experiments failed to yield any significant results, with the exception of the unexplained greater sensitivity in the monomodally-trained participants compared to the bimodally-trained ones on Day 3 of Experiment C1. It is unclear whether much should be read into this overall lack of significance. One could argue that participants simply need more exposure, or that the experiment was flawed in some other way. These experiments report on a “gap” between discrimination and word learning, but in reality, it may be premature to report this when the “end” of the gap was not found. That is, it is unknown how much training participants would need or what other factors would be necessary for participants to make use of their knowledge in a word learning task. The next step for these “C” Experiments may be to re-work these experiments to be more similar to those presented by Perfors and Dunbar (2010), Rost and McMurray (2009, 2010), or Fennell and Waxman (2010), who did find evidence that learners were able to extend the results of distributional learning to a word learning task. Although some weaknesses of Perfors and Dunbar (2010) were pointed out in Chapter 5 (namely, the unnaturalness of the task), it would be interesting to work from this study and gradually make the task more natural (for example, by adding in fillers) to see what parameters were necessary in order to observe this extension of distributional learning to word learning in an artificial language learning study.

#### 4.3. HOW “LINGUISTIC” IS DISTRIBUTIONAL LEARNING?

One point that I believe is important to keep in mind has to do with the nature of distributional learning. Maye and Gerken (2000) first introduce distributional learning as an alternative explanation to acquiring phonetic categories through minimal pairs. This can be argued since 1) infants exhibit language-specific discrimination of phones before learning minimal pairs (see Caselli et al., 1995), and 2) infants with the ability to discriminate two sounds ignore this same phonetic detail when presented with minimal pairs (Stager and Werker, 1997; Pater et al., 2004). However, the gap seen between distributional learning and the use of this information in a linguistically-meaningful way (i.e. to distinguish meaning) raises the question of whether distributional learning truly is “linguistic.” It may be the case that the presentation of minimal pairs or near-minimal pairs *is* necessary to form linguistically-relevant phonemes, where replacing one with the other causes a change in word meaning.

Seidl and Cristia (2012) distinguish between a “strong” form of this lexical hypothesis, that minimal pairs are necessary for sound acquisition, and a “weak” form of this hypothesis, lexical bootstrapping in the form of near-minimal pairs. The minimal pair hypothesis in its strong form is unlikely to be true for the reasons stated above. However, a lexical bootstrapping hypothesis does appear to be supported (Feldman et al., 2013; Thiessen, 2007; Swingley, 2009). Thiessen (2007) finds that 14-month olds fail to discriminate syllables beginning with [t] and [d] if trained on the minimal pairs [tɔ] and [dɔ], but *are* able to discriminate syllables beginning with [t] and [d] if trained on labels which differed in their overall wordforms, *dawbow* and *tawgoo*<sup>30</sup>. Being presented with clear evidence in the form of minimal pairs did not appear to indicate to the infant that [t] and [d] were contrastive. Rather, it was the presentation of distinct wordforms that brought about the increased discrimination between [t] and [d]. It may be the case that “distributional learning” results in early discriminatory abilities, but that these abilities are either so small as to not have a noticeable impact on word-learning, or are not utilized in linguistically-meaningful

---

<sup>30</sup> IPA transcriptions were not provided.

ways. That is, it is possible that distributional learning is not a mechanism for language acquisition, and instead is simply the residual side effect of talkers having some target gesture in mind. Another possibility is that distributional learning plays more of a role in shifting existing phoneme categories to accommodate talker variation, rather than in the acquisition of new categories (Xie et al., 2017).

#### 4.4. PHONETIC DISTANCE

A further question that has not been systematically studied in this dissertation has to do with phonetic distance between critical tokens. Maye and Gerken (2000) specifically test participants on continuum points 1 and 8, since both the bimodal and monomodal participants heard the exact same number of tokens of these two endpoints. However, it remains an open matter whether *perceptually* that remains the case. Endpoints used for these distributional learning experiments are by necessity very similar to one another, so that a smooth continuum can be synthesized. Although the endpoints, I believe, are perceptually distinct, consecutive (or even near-consecutive) continuum points are not. Any two consecutive continuum points are so similar that they may as well be identical from the participant's point of view. If this is the case, then it may very well be that bimodally-trained participants are *perceiving* Phone 1 and Phone 8 more than the monomodally-trained participants during training, and so during testing respond "different" more often due to the frequency of these sounds, rather than any sort of distributional inference. This study did not investigate this question, but I believe it is an important one for future distributional learning studies to keep in mind. It may be necessary to have a pre-test using the continuum points to ensure that each continuum point is at least 1 Just-Noticeable Difference (JND) away from its neighbor.

#### 4.5. WEAKNESSES AND FUTURE RESEARCH

##### 4.5.1 *L1 vs. L2*

One of the major weaknesses of the experiments presented here is that they seek to provide answers regarding early sound acquisition, but do so by conducting experiments on adults. One could argue that these experiments are more indicative of second language acquisition than first language acquisition.

However, these experiments also come up short when explaining second language acquisition, due to the artificial nature of the tasks presented to participants. Learners acquiring a second language likely receive more explicit instruction, whether through formal education or through interactions with native speakers in which minimal pairs are presented to illustrate a mistake that the learner is making.

Where possible, this study tried to draw parallels between experimental results and first language studies. However, the experiments presented here would undoubtedly benefit greatly by being followed up on work with infants. Specifically, I believe Experiment B regarding the timeline of allophony acquisition would greatly benefit from further infant studies.

#### 4.5.2 *Properties of the Stimuli*

Two types of critical continua were used in this dissertation: a stop continuum and a fricative continuum. These stimuli differed from one another in a number of ways. First, the two differ acoustically, with stops being abrupt by nature and fricatives being uniform and continuous throughout the duration of their production. (See the discussion in Chapter 3, end of Section 7). Second, the stop continuum used in this dissertation ranged between two allophones which are prototypical for English speakers: the prevoiced [g] is a prototypical allophone of word-initial /g/, and [g̊] is a prototypical allophone of /k/ when in an onset preceded by /s/. However, the fricative continuum ranged between two non-prototypical endpoints [ɛ] and [ɛ̃], which, to an English speaker are likely both perceived as non-prototypical variants of /f/.

Future work would greatly benefit from testing various properties of stimuli used and obtaining perceptual information. Ideally, I believe the following should be obtained for all critical stimuli to be used in distributional experiments: 1) perceptual distance between each critical continuum point, 2) prototypicality ratings to existing categories, 3) similarity judgments of the endpoints to one another. Regarding (2), it would be interesting to more systematically test the distributional learning of contrasts according to how the endpoints of the critical continuum are classified within the Perceptual Assimilation Model (Best, 1995).

#### 4.6. SUMMARY AND FUTURE STUDY

This dissertation has presented a series of artificial language learning experiments with the goal of outlining a timeline of early phonological acquisition in naïve adult learners. The two-stage model of distributional learning presented here is a domain-general model drawing on phenomena observed within psychology (e.g. across-category expansion and within-category compression), rather than one which is language-specific (e.g. the perceptual warping model presented by Boersma et al. (2003)). A few suggestions have been made throughout regarding areas for future research. Some of the most interesting topics for future research (I believe) lie in the following areas:

- 1) **Stimuli**: How do learners treat different types of stimuli in distributional learning? Could the typology of contrast types in Best's Perceptual Assimilation Model be incorporated into models of distributional learning?
- 2) **Allophony**: As far as I am aware, this dissertation and Noguchi (2016) are the only studies which have successfully shown that a decrease in sensitivity can be found in the lab by exposing learners to phones which are in complementary distribution, and both of these studies have been on adults. I believe this is an important area for more research, especially in the form of infant studies, utilizing a wider variety of critical stimuli.
- 3) **Attention**: I suggest in this dissertation that "distributional learning" may be the result of heightened awareness of variation in the speech stream. Further exploration of this idea may shed light on the underlying mechanism driving distributional learning.

And finally, I believe both past and future studies may benefit from re-analyzing their results in light of the two-stage model presented in this dissertation. As was shown in Chapter 3, the reported non-replication of Maye et al. (2002), Yoshida et al. (2011), appears to find support for a Bias Stage. I believe past and future work will benefit from considering distributional learning as a two-stage process.

**APPENDIX 3.1: NUMBER OF PARTICIPANTS IN “A” EXPERIMENTS**

	<b>A1</b>	<b>A2</b>	<b>A3</b>	<b>A2-Tone</b>	<b>A3-Tone</b>
<b>Exclusion reason</b>					
Sound check (< 5/6)	12	7	28	14	21
Practice (< 5/8)	0	1	4	0	8
Train Check tone-monitoring task (< 10/24) <sup>31</sup>	N/A	N/A	N/A	(2) <sup>32</sup>	Data not collected
Not native English speaker	0	0	1	0	2
Speech/hearing disorder	1	1	0	0	5
Experience with language with more than one voiceless post-alveolar fricative	N/A	N/A	3	N/A	5
<b>Total rejected</b> (some for multiple reasons)	13	7	32	14	28
<b>Total participants</b>					
Bimodal	27	21	22	28	24
Monomodal	34	27	27	31	19

---

<sup>31</sup> In actuality, the first experiment to be completed was a pilot “Tone” experiment which contained the Train Check tone-monitoring task. The original goal of the tone-monitoring task had been to ensure that participants were paying attention during the Train phase. Because of this, participants who correctly responded to fewer than 10/24 of monitoring trials were to be excluded from analysis. However, after it was hypothesized that the monitoring trials might have a negative effect on distributional learning, the “Non-tone” experiments (Experiments A1-A3) were designed. In order to directly compare the effect of including monitoring tokens on participants’ behavior, it was decided that participants should not be excluded for incorrect responses to monitoring trials, since only the “Tone” experiments contained these trials and comparisons of results could be attributed to testing different populations (those who were paying attention and therefore passed the monitoring task, compared to the general population). Therefore, this exclusion criterion was not utilized in this dissertation.

That being said, the number of participants who would have been excluded under this criterion is still given here for reference. It happens that these 2 participants who failed this criterion had also failed the sound check, and therefore would not have been included in the analysis even if the monitoring task criterion had been adhered to.

<sup>32</sup> As noted in the Footnote 31, these 2 participants also happened to fail the Sound check criterion, and therefore would not have been included in the analysis either way.

## APPENDIX 3.2: PARTICIPANT QUESTIONNAIRE FOR “A” EXPERIMENTS

The text that was given to participants during the questionnaire section of all “A” Experiments is shown below.

*[Page 1]*

*You're almost done! Please fill out the following questionnaire regarding your language background. You will NOT be negatively affected based on your answers, so please answer truthfully.*

*Press the “Y” to continue to the questionnaire*

*[Page 2]*

*Do you remember any of the words you heard from this artificial language? If so, please list them below.*

*If you noticed any patterns in the foreign language you heard, please list them.*

*Did you use a strategy when learning this artificial language or when trying to determine whether words were the same or different? If so, what strategy did you use?*

*[Page 3]*

*Where have you lived?*

*What languages do you have experience with? For example, list any languages your parents may speak, or any languages you have studied in school or have studied on your own. For each, please indicate how fluent you are (for example: how long you've studied it, how comfortable you feel with each language...).*

*[Page 3]*

*How comfortable are you with English?*

- I'm a native speaker*
- I'm pretty comfortable with speaking/understanding English*
- I am not comfortable with speaking/understanding English*



*How old are you?*

- 18-25 years old*
- 26-35 years old*
- 36-45 years old*
- 46-65 years old*
- Over 66 years old*
- Prefer not to answer*

*What is your sex?*

- Female*
- Male*
- Other/prefer not to answer*

*Do you have any background studying or reading about linguistics or phonetics?*

- Yes*
- No*
- Prefer not to respond*

*Do you have a history of a speech or hearing disorder?*

- Yes*
- No*
- Prefer not to respond*

*During the LISTENING portion (where you simply listened to words), how much attention would you say you were paying to this study? (Remember, your answer will NOT affect payment, so please answer truthfully!)*

- I focused all of my attention on this portion of the experiment*
- I mostly paid attention*
- I was not paying very much attention*
- I paid very little attention*

*During the LISTENING portion, were you wearing headphones the entire time?*

- Yes, I was wearing headphones the entire time*
- I was wearing headphones most of the time*
- I was wearing headphones some of the time*
- No, I did not wear headphones at all*

*During the TESTING portion (where you heard two words and were asked whether they were the same or different in this language), how much attention would you say you were paying to this study? (Remember, your answer will NOT affect payment, so please answer truthfully!)*

- I focused all of my attention on this portion of the experiment*
- I mostly paid attention*
- I was not paying very much attention*
- I paid very little attention*

*During the TESTING portion, were you wearing headphones the entire time?*

- Yes, I was wearing headphones the entire time*
- I was wearing headphones most of the time*
- I was wearing headphones some of the time*
- No, I did not wear headphones at all*

*If you were taking this same experiment in a lab setting, do you think you would pay...*

- More attention*
- Less attention*
- About the same level of attention*

*[Page 4]*

*Were you doing other things during the course of this experiment? Were there outside distractions? If so, please briefly explain.*

*If you have any other comments regarding this experiment, please write them here.*

**APPENDIX 4.1: NUMBER OF PARTICIPANTS IN EXPERIMENT B**

	<b>B</b>
<b>Exclusion reason</b>	
Sound check (not included to keep experiment short)	N/A
Practice (< 5/8)	15
Not native English speaker	1
Speech/hearing disorder	4
Experience with language with more than one voiceless post-alveolar fricative	22
<b>Total rejected</b> (some for multiple reasons)	40
ExposureTime One	15
ExposureTime Two	12
ExposureTime Three	13
<b>Total participants</b>	
Bimodal-Comp	
ExposureTime One	41 (23 PhoneFirst, 18 RuleFirst)
ExposureTime Two	40 (23 PhoneFirst, 17 RuleFirst)
ExposureTime Three	43 (25 PhoneFirst, 18 RuleFirst)
Bimodal-NonComp	
ExposureTime One	51 (31 PhoneFirst, 20 RuleFirst)
ExposureTime Two	33 (17 PhoneFirst, 16 RuleFirst)
ExposureTime Three	40 (19 PhoneFirst, 21 RuleFirst)
Monomodal	
ExposureTime One	48 (28 PhoneFirst, 20 RuleFirst)
ExposureTime Two	44 (23 PhoneFirst, 21 RuleFirst)
ExposureTime Three	51 (26 PhoneFirst, 25 RuleFirst)

## APPENDIX 4.2: PARTICIPANT QUESTIONNAIRE FOR EXPERIMENT B

The text that was given to participants during the questionnaire section of Experiment B is shown below.

*[Page 1]*

*You're almost done! Please fill out the following questionnaire regarding your language background. You will NOT be negatively affected based on your answers, so please answer truthfully.*

*Press the “Y” to continue to the questionnaire*

*[Page 2]*

*Do you remember any of the words you heard from this artificial language? If so, please list them below.*

*If you noticed any patterns in the foreign language you heard, please list them.*

*How did you approach the tasks in this experiment? Did you use a strategy when learning this artificial language or during either test phase? If so, what strategy did you use?*

*[Page 3]*

*Where have you lived for more than 3 years? Please provide state and country.*

*What languages do you have experience with? For example, list any languages your parents may speak, or any languages you have studied in school or have studied on your own. For each, please indicate how fluent you are (for example: how long you've studied it, how comfortable you feel with each language...).*

*Have you studied linguistics or phonetics? If so, please briefly explain here.*

*[Page 3]*

*How comfortable are you with English?*

- I'm a native speaker*
- I'm pretty comfortable with speaking/understanding English*
- I am not comfortable with speaking/understanding English*

*How old are you?*

- 18-25 years old*
- 26-35 years old*
- 36-45 years old*
- 46-65 years old*
- Over 66 years old*
- Prefer not to answer*

*What is your sex?*

- Female*
- Male*
- Other/prefer not to answer*

*Do you have a history of a speech or hearing disorder?*

- Yes*
- No*
- Prefer not to respond*

*Please be honest (your answer will NOT affect your payment) -- how much were you listening during this experiment?*

- I paid attention the whole time, and I listened to the whole experiment*
- I mostly paid attention, and I listened to the whole experiment*
- I did not pay very much attention, but I listened to the whole experiment*
- I listened to most of the experiment*
- I only listened to part of the experiment*

*[Page 4]*

*Please specify your attention level during this experiment. For example, were there any distractions in the room you were taking this in? Did you remove your headphones? If so, for how long?*

**APPENDIX 5.1: NUMBER OF PARTICIPANTS IN “C” EXPERIMENTS**

	<b>C1</b>	<b>C2</b>
<b>Exclusion reason</b>		
Sound check (< 15/18)	11	12
Practice (< 15/24)	0	0
Not native English speaker	1	0
Speech/hearing disorder	3	0
Experience with language with more than one voiceless post-alveolar fricative	N/A	10
Reported writing down words during experiment	1	0
<b>Total rejected</b> (some for multiple reasons)	15	21
<b>Total participants</b>		
Total participated at least one day	69	104
Total participated all three days	52	68
Bimodal	18	27
Monomodal	19	20

## APPENDIX 5.2: PARTICIPANT QUESTIONNAIRE FOR “C” EXPERIMENTS

The text that was given to participants during the questionnaire section of all “C” Experiments is shown below.

*[Page 1]*

*You're almost done! Please fill out the following questionnaire regarding your language background. You will NOT be negatively affected based on your answers, so please answer truthfully.*

*Press the “Y” to continue to the questionnaire*

*[Page 2]*

*Do you remember any of the words you heard from this artificial language? If so, please list them below.*

*If you noticed any patterns in the foreign language you heard, please list them.*

*Did you use a strategy when learning this artificial language or when trying to determine whether words were the same or different? If so, what strategy did you use?*

*[Page 3]*

*Where have you lived?*

*What languages do you have experience with? For example, list any languages your parents may speak, or any languages you have studied in school or have studied on your own. For each, please indicate how fluent you are (for example: how long you've studied it, how comfortable you feel with each language...).*

*[Page 3]*

*How comfortable are you with English?*

- I'm a native speaker*
- I'm pretty comfortable with speaking/understanding English*
- I am not comfortable with speaking/understanding English*

*What is your sex?*

- Female*
- Male*
- Other/prefer not to answer*

*Do you have any background studying or reading about linguistics or phonetics?*

- Yes*
- No*
- Prefer not to respond*

*Do you have a history of a speech or hearing disorder?*

- Yes*
- No*
- Prefer not to respond*

*How much attention would you say you were paying to this study on Day 1? (Remember, your answer will NOT affect payment, so please answer truthfully!)*

- I focused all of my attention on this portion of the experiment*
- I mostly paid attention*
- I was not paying very much attention*
- I paid very little attention*

*How much attention would you say you were paying to this study on Day 2? (Remember, your answer will NOT affect payment, so please answer truthfully!)*

- I focused all of my attention on this portion of the experiment*
- I mostly paid attention*
- I was not paying very much attention*
- I paid very little attention*

*How much attention would you say you were paying to this study on Day 3? (Remember, your answer will NOT affect payment, so please answer truthfully!)*

- I focused all of my attention on this portion of the experiment*
- I mostly paid attention*



- *I was not paying very much attention*
- *I paid very little attention*

*Were you wearing headphones the entire time on Day 1?*

- *Yes, I was wearing headphones the entire time*
- *I was wearing headphones most of the time*
- *I was wearing headphones some of the time*
- *No, I did not wear headphones at all*

*Were you wearing headphones the entire time on Day 2?*

- *Yes, I was wearing headphones the entire time*
- *I was wearing headphones most of the time*
- *I was wearing headphones some of the time*
- *No, I did not wear headphones at all*

*Were you wearing headphones the entire time on Day 3?*

- *Yes, I was wearing headphones the entire time*
- *I was wearing headphones most of the time*
- *I was wearing headphones some of the time*
- *No, I did not wear headphones at all*

*If you were taking this same experiment in a lab setting, do you think you would pay...*

- *More attention*
- *Less attention*
- *About the same level of attention*

*[Page 4, only included in Experiment C2]*

*In this experiment, there were two 'sh'-like sounds. Did you notice that there were two?*

*If you noticed that there were two 'sh'-like sounds, did you think that switching one of these 'sh' sounds for the other caused a change in word meaning?*

*For example, in English we can say 'help' with more of a closed 'eh' sound ('h(eh)lp') or with more of an open 'ah' sound ('h(ah)lp'), but these different pronunciations still refer to the same word. But if we switch the 'h' sound with a 'k' sound, it now refers to a type of seaweed, 'kelp', and so DOES cause a change in word meaning.*

*Did you think switching one of these 'sh' sounds for the other 'sh' sound did NOT cause a change in word meaning (as in 'help'~'halp'), or did you think that switching one for the other DID cause a change in word meaning (as in 'help'~'kelp')?*

*Is there anything you noticed about the 'sh' sound(s) while taking this experiment?*

*[Page 4 (Experiment C1) or 5 (Experiment C2)]*

*PLEASE BE HONEST: Did you write down any of the words from the artificial language? If so, which words?*

*Were you doing other things during the course of this experiment? Were there outside distractions? If so, please briefly explain.*

*If you have any other comments regarding this experiment, please write them here.*

## REFERENCES

- Adriaans, Frans, and Daniel Swingley. 2012. "Distributional learning of vowel categories is supported by prosody in infant-directed speech." *Annual Meeting of the Cognitive Science Society*.
- Aslin, Richard N, and David B Pisoni. 1980. "Effects of Early Linguistic Experience on Speech Discrimination by Infants: A Critique of Eilers, Gavin, and Wilson (1979)." *Child Development* 51 (1): 107.
- Aslin, Richard N, and David B Pisoni. 1980. "Some developmental processes in speech perception." *Child Phonology* 2.
- Baddeley, A.D, and W.P Colquhoun. 1969. "Signal probability and vigilance: A reappraisal of the 'signal-rate' effect." *British Journal of Psychology* 60 (2): 169.
- Barr, Dale J. 2013. "Random Effects Structure for Testing Interactions in Linear Mixed-Effects Models." *Frontiers in Psychology* 4: 328.
- Barr, Dale J, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. "Random effects structure for confirmatory hypothesis testing: Keep it maximal." *Journal of Memory and Language* 68 (3): 255-278.
- Bates, Douglas, Martin Maechler, Ben Bolker, and Steven Walker. 2015. "lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7. 2014."
- Beale, James M, and Frank C Keil. 1995. "Categorical effects in the perception of faces." *Cognition* 57 (3): 217-239.
- Benesty, Jacob, M. Mohan Sondhi, and Yiteng Arden Huang. 2008. *Springer Handbook of Speech Processing*. Berlin: Springer.
- Bergelson, Erika, and Daniel Swingley. 2015. "Early word comprehension in infants: Replication and extension." *Language Learning and Development* 11 (4): 369-380.
- Best, Catherine T. 1994. "The emergence of native-language phonological influences in infants: A perceptual assimilation mode." *The development of speech perception: The transition from speech sounds to spoken word* 167 (224): 233-277.
- Best, Catherine T, Gerald W McRoberts, and Nomathemba M Sithole. 1988. "Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants." *Journal of Experimental Psychology: Human Perception and Performance* 14 (3): 345.
- Best, Catherine T, Gerald W McRoberts, Rosemarie LaFleur, and Jean Silver-Isenstadt. 1995. "Divergent developmental patterns for infants' perception of two nonnative consonant contrasts." *Infant Behavior and Development* 18 (3): 339-350.
- Best, Catherine, and Gerald McRoberts. 1989. "Phonological influence on infants' perception of two nonnative speech contrasts."

- Bion, Ricardo AH, Kouki Miyazawa, Hideaki Kikuchi, and Reiko Mazuka. 2013. "Learning phonemic vowel length from naturalistic recordings of Japanese infant-directed speech." *PLOS ONE* 8 (2): e51594.
- Blomert, Leo, and Holger Mitterer. 2004. "The fragile nature of the speech-perception deficit in dyslexia: Natural vs. synthetic speech." *Brain and Language* 89 (1): 21-26.
- Boersma, Paul. 1998. "Functional Phonology: Formalizing the interactions between articulatory and perceptual drives." *The Hague: Holland Academic Graphics* 11.
- Boersma, Paul, and Bruce Hayes. 2001. "Empirical tests of the gradual learning algorithm." *Linguistic Inquiry* 32 (1): 45-86.
- Boersma, Paul, Paola Escudero Escudero, and Rachel Hayes. 2003. "Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories." *15th International Congress of Phonetic Sciences*. 1013-1016.
- Boomershine, Amanda, Kathleen Currie Hall, Elizabeth Hume, and Keith Johnson. 2008. "The impact of allophony versus contrast on speech perception." In *Contrasts in phonology: Theory, perception, acquisition*, 145-171.
- Brent, Michael R, and Jeffrey M Siskind. 2001. "The role of exposure to isolated words in early vocabulary development." *Cognition* 81 (2): B33-B44.
- Burns, Edward M, and W. Dixon Ward. 1978. "Categorical perception—phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals." *The Journal of the Acoustical Society of America* 63 (2): 456-468.
- Caselli, Maria Cristina, Elizabeth Bates, Paola Casadio, Judi Fenson, Larry Fenson, Lisa Sanderl, and Judy Weir. 1995. "A cross-linguistic study of early lexical development." *Cognitive Development* 10 (2): 159-199.
- Cheour, Marie, Rita Ceponiene, Anne Lehtokoski, Aavo Luuk, Jüri Allik, Kimmo Alho, and Risto Näätänen. 1998. "Development of language-specific phoneme representations in the infant brain." *Nature Neuroscience* 1 (5): 351.
- Clark, Herbert H. 1973. "The language-as-fixed-effect fallacy: A critique of language statistics in psychological research." *Journal of Verbal Learning and Verbal Behavior* 12 (4): 335-359.
- Cristia, Alejandrina. 2018. "Can infants learn phonology in the lab? A meta-analytic answer." *Cognition* 170: 312-327.
- Cristia, Alejandrina, Grant L McGuire, Amanda Seidl, and Alexander L Francis. 2011. "Effects of the distribution of acoustic cues on infants' perception of sibilants." *Journal of Phonetics* 39 (3): 388-402.
- Crump, Matthew, John V McDonnell, and Todd M Gureckis. 2013. "Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research." *PLOS One* 8 (3): e57410.
- Cutler, Anne, Jacques Mehler, Dennis Norris, and Juan Segui. 1987. "Phoneme identification and the lexicon."

- Daidone, Danielle, and Isabelle Darcy. 2014. "Quiero comprar una guitarra: Lexical Encoding of the Tap and Trill by L2 Learners of Spanish." *Selected Proceedings of the Second Language Research Forum*. Somerville: Cascadilla. 30-38.
- Darcy, Isabelle, Danielle Daidone, and Chisato Kojima. 2013. "Asymmetric lexical access and fuzzy lexical representations in second language learners." *The Mental Lexicon* 8 (3): 372-420.
- Davenport, W. G. 1969. "Vibrotactile vigilance: The effects of costs and values on signals." *Perception & Psychophysics* 5 (1): 25-28.
- de Beeck, Hans Op, Johan Wagemans, and Rufin Vogels. 2003. "The effect of category learning on the representation of shape: Dimensions can be biased but not differentiated." *Journal of Experimental Psychology: General* 132 (4): 491.
- DeCarlo, Lawrence T. 2013. "Signal detection models for the same-different task." *Journal of Mathematical Psychology* 57: 43-51.
- DeCarlo, Lawrence T. 1998. "Signal detection theory and generalized linear models." *Psychological Methods* 3 (2): 186.
- DeCasper, Anthony J, and Melanie J Spence. 1986. "Prenatal maternal speech influences newborns' perception of speech sounds." *Infant Behavior and Development* 9 (2): 133-150.
- DeCasper, Anthony J, and William P Fifer. 1980. "Of human bonding: Newborns prefer their mothers' voices." *Science* 208 (4448): 1174-1176.
- DeCasper, Anthony J, Jean-Pierre Lecanuet, Marie-Claire Busnel, Carolyn Granier-Deferre, and Roselyne Maugeais. 1994. "Fetal reactions to recurrent maternal speech." *Infant Behavior and Development* 17 (2): 159-164.
- Dehaene-Lambertz, Ghislaine, Stanislas Dehaene, and Lucie Hertz-Pannier. 2002. "Functional neuroimaging of speech perception in infants." *Science* 298 (5600): 2013-2015.
- Denby, Thomas, Jeffrey Schecter, Sean Arn, Svetlin Dimov, and Matthew Goldrick. 2017. "Contextual variability and exemplar strength in phonotactic learning." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 44 (2): 280-294.
- Dillon, Brian, Ewan Dunbar, and William Idsardi. 2013. "A Single-Stage Approach to Learning Phonological Categories: Insights From Inuktitut." *Cognitive Science* 1: 34.
- Dougal, Sonya, and Caren M Rotello. 2007. "'Remembering' emotional words is based on response bias, not recollection." *Psychonomic Bulletin & Review* 14 (3): 423-429.
- Dumay, Nicolas, and M. Gareth Gaskell. 2007. "Sleep-associated changes in the mental representation of spoken words." *Psychological Science* 18 (1): 35-39.
- Eilers, Rebecca E, William Gavin, and Wesley R Wilson. 1979. "Linguistic experience and phonemic perception in infancy: A crosslinguistic study." *Child Development* 14-18.
- Eilers, Rebecca E, William J Gavin, and Wesley R Wilson. 1980. "Effects of early linguistic experience on speech discrimination by infants: A reply." *Child Development* 113-117.

- Eimas, Peter D. 1978. "Developmental aspects of speech perception." In *Perception*, by Richard Held, Herschel W Leibowitz and Hans-Lukas Teuber, 357-374. Berlin, Heidelberg: Springer.
- Eimas, Peter D, Siqueland, Peter Jusczyk, and James Vigorito. 1971. "Speech perception in infants." *Science* 171 (3968 ): 303-306.
- Escudero, Paola, and Daniel Williams. 2014. "Distributional learning has immediate and long-lasting effects." *Cognition* 133 (2): 408-413.
- Escudero, Paola, Titia Benders Benders, and Karin Wanrooij. 2011. "Enhanced bimodal distributions facilitate the learning of second language vowels." *The Journal of the Acoustical Society of America* 130 (4): EL206-EL212.
- Feldman, Naomi H, Emily B Myers, Katherine S White, Thomas L Griffiths, and James L Morgan. 2013. "Word-level information influences phonetic learning in adults and infants." *Cognition* 127 (3): 427-438.
- Feldman, Naomi, Emily Myers, Katherine White, Thomas Griffiths Griffiths, and James L Morgan. 2011. "Learners use word-level statistics in phonetic category acquisition." *Proceedings of the 35th Boston University Conference on Language Development*. 197-209.
- Feldman, Naomi, Thomas Griffiths Griffiths, and James Morgan. 2009. "Learning phonetic categories by learning a lexicon." *Cognitive Science Society*.
- Fenn, Kimberly M, Howard C Nusbaum, and Daniel Margoliash. 2003. "Consolidation during sleep of perceptual learning of spoken language." *Nature* 425 (6958): 614.
- Fennell, Christopher T., and Sandra R. Waxman. 2010. "What paradox? Referential cues allow for infant use of phonetic detail in word learning." *Child Development* 81 (5): 1376-1383. Accessed 11 15, 2018. <http://psychology.northwestern.edu/documents/waxman-paradox.pdf>.
- Ferguson, Brock, Mélanie Havy, and Sandra R Waxman. 2015. "The precision of 12-month-old infants' link between language and categorization predicts vocabulary size at 12 and 18 months." *Frontiers in Psychology* 1319.
- Galle, Marcus E, Keith S Apfelbaum, and Bob McMurray. 2015. "The role of single talker acoustic variation in early word learning." *Language Learning and Development* 11 (1): 66-79.
- Ganger, Jennifer, and Michael R Brent. 2004. "Reexamining the vocabulary spurt." *Developmental Psychology* 40 (4): 621.
- Gaskell, M. Gareth, and Nicolas Dumay. 2003. "Lexical competition and the acquisition of novel words." *Cognition* 89 (2): 105-132.
- Gervain, Judit, and Jacques Mehler. 2010. "Speech perception and language acquisition in the first year of life." *Annual Review of Psychology* 61: 191-218.
- Gervain, Judit, and Janet F Werker. 2008. "How infant speech perception contributes to language acquisition." *Language and Linguistics Compass* 2 (6): 1149-1170.

- Goldfield, Beverly A, and J. Steven Reznick. 1990. "Early lexical acquisition: Rate, content, and the vocabulary spurt." *Journal of Child Language* 17 (1): 171-183.
- Goldstone, Robert L. 1994. "Influences of categorization on perceptual discrimination." *Journal of Experimental Psychology: General* 123 (2): 178.
- Goldstone, Robert L, and Andrew T Hendrickson. 2010. "Categorical perception." *Wiley Interdisciplinary Reviews: Cognitive Science* 1 (1): 69-78.
- Gosling, Samuel D, Simine Vazire, Sanjay Srivastava, and John P Oliver. 2004. "Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires." *American Psychologist* 59 (2): 93.
- Guenther, Frank H, and Marin N Gjaja. 1996. "The perceptual magnet effect as an emergent property of neural map formation." *The Journal of the Acoustical Society of America* 100 (2): 1111-1121.
- Guenther, Frank H, Fatima T Husain, Michael A Cohen, and Barbara G Shinn-Cunningham. 1999. "Effects of categorization and discrimination training on auditory perceptual space." *The Journal of the Acoustical Society of America* 106 (5): 2900-2912.
- Guion, Susan G, James E Flege, Reiko Akahane-Yamada, and Jesica C Pruitt. 2000. "An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants." *The Journal of the Acoustical Society of America* 107 (5): 2711-2724.
- Gulian, Margarita, Escudero Paola, and Paul Boersma. 2007. "Supervision hampers distributional learning of vowel contrasts." *16th International Congress of Phonetic Sciences*. Saarbrücken, Germany: Saarland University. 1893-1896.
- Hallé, Pierre A, and Bénédicte de Boysson-Bardies. 1996. "The format of representation of recognized words in infants' early receptive lexicon." *Infant Behavior and Development* 19 (4): 463-481.
- Harris, Zellig Sabbetai. 1963. *Structural linguistics*. 6th. Chicago: The University of Chicago Press.
- Hayes-Harb, Rachel. 2007. "Lexical and statistical evidence in the acquisition of second language phonemes." *Second Language Research* 23 (1): 65-94.
- Hothorn, Torsten, Frank Bretz, Peter Westfall, Richard M Heiberger, Andre Schuetzenmeister, and Susan Scheibe. 2017. *Package 'multcomp'*.
- Iverson, Paul, and Patricia K Kuhl. 1995. "Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling." *The Journal of the Acoustical Society of America* 97 (1): 553-562.
- Iverson, Paul, Patricia K Kuhl, Reiko Akahane-Yamada, Eugen Diesch, Yoh'ich Tohkura, Andreas Kettermann, and Claudia Siebert. 2003. "A perceptual interference account of acquisition difficulties for non-native phonemes." *Cognition* 87 (1): B47-B57.
- Jaeger, T. Florian. 2008. "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models." *Journal of memory and language* 59 (4): 434-446.

- Johnson, Keith, and Molly Babel. 2010. "On the perceptual basis of distinctive features: Evidence from the perception of fricatives by Dutch and English speakers." *Journal of Phonetics* 38 (1): 127-136.
- Jongman, Allard, Yue Wang, and Brian H Kim. 2003. "Contributions of semantic and facial information to perception of nonsibilant fricatives." *Journal of Speech, Language, and Hearing Research* 46 (6): 1367-1377.
- Kawahara, Hideki, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Toshio Irino, and Hideki Banno. 2008. "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation." *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. Las Vegas: IEEE. 3933-3936.
- Kay-Raining Bird, Elizabeth, and Robin S Chapman. 1998. "Partial representations and phonological selectivity in the comprehension of 13-to 16-month-olds." *First Language* 18 (52): 105-127.
- Kleinschmidt, Dave F, and T. Florian Jaeger. 2012. "A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation." *Annual Meeting of the Cognitive Science Society*.
- Kleinschmidt, Dave F, and T. Florian Jaeger. 2015. "Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel." *Psychological Review* 122 (2): 148.
- Kleinschmidt, Dave. 2017. *Perception in a variable but structured world: the case of speech perception*. Doctoral dissertation, Rochester, NY: University of Rochester.
- Kuhl, Patricia K. 1981. "Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories." *The Journal of the Acoustical Society of America* 70 (2): 340-349.
- Kuhl, Patricia K. 2004. "Early language acquisition: cracking the speech code." *Nature Reviews Neuroscience* 5 (11): 831-843.
- Kuhl, Patricia K, and Denise M Padden. 1983. "Enhanced discriminability at the phonetic boundaries for the place feature in macaques." *The Journal of the Acoustical Society of America* 73 (3): 1003-1010.
- Kuhl, Patricia K, and Denise M Padden. 1982. "Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques." *Perception & Psychophysics* 32 (6): 542-550.
- Kuhl, Patricia K, and James D Miller. 1978. "Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli." *The Journal of the Acoustical Society of America* 63 (3): 905-917.
- Kuhl, Patricia K, and James D Miller. 1975. "Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants." *Science* 190 (4209): 69-72.
- Kuhl, Patricia K, Erica Stevens, Akiko Hayashi, Toshisada Deguchi, Shigeru Kiritani, and Paul Iverson. 2006. "Infants show a facilitation effect for native language phonetic perception between 6 and 12 months." *Developmental Science* 9 (2).



- Kuhl, Patricia K, Karen A Williams, Francisco Lacerda, Kenneth N Stevens, and Björn Lindblom. 1992. "Linguistic experience alters phonetic perception in infants by 6 months of age." *Science* 255 (5044 ): 606-608.
- Kurtz, Kenneth J, and Dedre Gentner. 1998. "Category learning and comparison in the evolution of similarity structure." *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*.
- Lalonde, Chris E, and Janet F Werker. 1995. "Cognitive influences on cross-language speech perception in infancy." *Infant Behavior and Development* 18 (4): 459-475.
- Leach, Laura, and Arthur G Samuel. 2007. "Lexical configuration and lexical engagement: When adults learn new words." *Cognitive Psychology* 55 (4): 306-353.
- Lenneberg, Eric H. 1967. "The biological foundations of language." *Hospital Practice* 2 (12): 59-67.
- Levin, Daniel T, and James M Beale. 2000. "Categorical perception occurs in newly learned faces, other-race faces, and inverted faces." *Perception & Psychophysics* 62 (2): 386-401.
- Lisker, Leigh, and Arthur S Abramson. 1964. "A cross-language study of voicing in initial stops: Acoustical measurements." *Word* 20 (3): 384-422.
- Liu, Liquan, and René Kager. 2014. "Perception of tones by infants learning a non-tone language." *Cognition* 133 (2): 385-394.
- Livingston, Kenneth R, Janet K Andrews, and Stevan Harnad. 1998. "Categorical perception effects induced by category learning." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24 (3): 732.
- Luce, R. Duncan. 1959. *Individual Choice Behavior*. Vol. 70. New York: Wiley.
- MacKain, Kristine S. 1982. "Assessing the role of experience on infants' speech discrimination." *Journal of Child Language* 9 (3): 527-542.
- MacKain, Kristine S, and Daniel N Stern. 1985. *The concept of experience in speech development*. Vol. 5, in *Children's Language*, by Keith E Nelson, 1-33. New York: Psychology Press.
- Macmillan, Neil A, and C. Douglas Creelman. 2004. *Detection theory: A user's guide*. 2nd. Mahwah, NJ: Lawrence Erlbaum Associates.
- Mason, Winter, and Siddharth Suri. 2012. "Conducting behavioral research on Amazon's Mechanical Turk." *Behavior Research Methods* 44 (1): 1-23.
- Mattock, Karen, and Denis Burnham. 2006. "Chinese and English infants' tone perception: Evidence for perceptual reorganization." *Infancy* 10 (3): 241-265.
- Mattock, Karen, Monika Molnar, Linda Polka, and Denis Burnham. 2008. "The developmental course of lexical tone perception in the first year of life." *Cognition* 106 (3): 1367-1381.
- Maye, Jessica, and LouAnn Gerken. 2000. "Learning phonemes without minimal pairs." *24th Annual Boston University Conference on Language Development*.

- . 2001. "Learning phonemes: How far can the input take us." *25th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press. 480.
- Maye, Jessica, Daniel J Weiss, and Richard N Aslin. 2008. "Statistical phonetic learning in infants: Facilitation and feature generalization." *Developmental Science* 1 (1): 122-134.
- Maye, Jessica, Janet F Werker, and LouAnn Gerken. 2002. "Infant sensitivity to distributional information can affect phonetic discrimination." *Cognition* 82 (3): B101-B111.
- McAuliffe, Michael, and Molly Babel. 2016. "Stimulus-directed attention attenuates lexically-guided perceptual learning." *The Journal of the Acoustical Society of America* 140 (3): 1727-1738.
- McGuire, Grant L. 2007. *Phonetic category learning*. Doctoral dissertation, The Ohio State University.
- Mehler, Jacques, Josiane Bertoncini, Michele Barriere, and Dora Jassik-Gerschenfeld. 1978. "Infant recognition of mother's voice." *Perception* 7 (5): 491-497.
- Mehler, Jacques, Peter Jusczyk, Ghislaine Lambertz, Nilofar Halsted, Josiane Bertoncini, and Claudine Amiel-Tison. 1988. "A precursor of language acquisition in young infants." *Cognition* 29 (2): 143-178.
- Minai, Utako, Kathleen Gustafson, Robert Fiorentino, Allard Jongman, and Joan Sereno. 2017. "Fetal rhythm-based language discrimination: a biomagnetometry study." *NeuroReport* 28 (10): 561-564.
- Moeng, Emily. 2016. "Comparing the distributional learnability of stops, fricatives, glides, and vowels." *40th Boston University Conference on Language Development*. Boston, MA.
- . 2017. "Distributional learning on Mechanical Turk and effects of attentional shifts." *Proceedings of the Linguistic Society of America*. 1-15. doi:<http://dx.doi.org/10.3765/plsa.v2i0.4105>.
- Moreton, Elliott, and Joe Pater. 2012. "Structure and substance in artificial-phonology learning, part II: Substance." *Language and Linguistics Compass* 6 (11): 702-718.
- Morse, Philip A. 1978. "Infant speech perception: Origins, processes and alpha centauri." In *Communicative and Cognitive Abilities: Early Behavioral Assessment*. Baltimore: University Park Press.
- Narayan, Chandan R, Janet F Werker, and Patrice Speeter Beddor. 2010. "The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination." *Developmental Science* 13 (3): 407-420.
- Noguchi, Masaki. 2016. *Acquisition of allophony from speech input by adult learners*. Doctoral dissertation, Vancouver: The University of British Columbia.
- Nosofsky, Robert M. 1986. "Attention, similarity, and the identification–categorization relationship." *Journal of Experimental Psychology: General* 115 (1): 39.
- Ong, Jia Hoong, Denis Burnham, and Paola Escudero. 2015. "Distributional learning of lexical tones: A comparison of attended vs. unattended listening." *PLOS One* 10 (7): e0133446.

- Ong, Jia, Josephine Terry, and Paola Escudero. 2016. "Can Australian English listeners learn non-native vowels via distributional learning?" *Sixteenth Australasian International Conference on Speech Science and Technolog.* Parramatta, Australia. 289-292.
- Özgen, Emre, and Ian RL Davies. 2002. "Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis." *Journal of Experimental Psychology: General* 141 (4): 477.
- Pajak, Bożena, and Roger Levy. 2012. "Distributional learning of L2 phonological categories by listeners with different language backgrounds." *Proceedings of the 36th Boston University Conference on Language Development.* Somerville, MA: Cascadilla Press. 400-413.
- . 2011. "Phonological generalization from distributional evidence." *Proceedings of the Annual Meeting of the Cognitive Science Society.* 2673-2678.
- Pajak, Bożena, Sarah C Creel, and Roger Levy. 2016. "Difficulty in learning similar-sounding words: A developmental stage or a general property of learning?" *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42 (9): 1377.
- Paolacci, Gabriele, and Jesse Chandler. 2014. "Inside the Turk: Understanding Mechanical Turk as a participant pool." *Directions in Psychological Science* 23 (3): 184-188.
- Pater, Joe, Christine Stager Stager, and Janet Feldman Werker. 2004. "The perceptual acquisition of phonological contrasts." *Language* 80 (3): 384-402.
- Pegg, Judith E, and Janet F Werker. 1997. "Adult and infant perception of two English phones." *The Journal of the Acoustical Society of America* 102 (6): 3742-3753.
- Pena, Marcela, Atsushi Maki, Damir Kovačić, Ghislaine Dehaene-Lambertz, Hideaki Koizumi, Furio Bouquet, and Jacques Mehler. 2003. "Sounds and silence: an optical topography study of language recognition at birth." *Proceedings of the National Academy of Sciences* 100 (20): 11702-11705.
- Peperkamp, Sharon, Michèle Pettinato Pettinato, and Emmanuel Dupoux. 2003. "Allophonic variation and the acquisition of phoneme categories." *27th Annual Boston University Conference on Language Development.* Boston, MA: Cascadilla Press. 650-661.
- Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. 2006. "The acquisition of allophonic rules: Statistical learning with linguistic constraints." *Cognition* 101 (3): B31-B41.
- Perfors, Amy, and David Dunbar. 2010. "Phonetic training makes word learning easier." *Cognitive Science Society.*
- Peterson, Wesley W, Theodore G Birdsall, and W.C Fox. 1954. "The theory of signal detectability." *Transactions of the IRE Professional Group on Information Theory* 4 (4): 171-212.
- Pitt, Mark A, and Christine M Szostak. 2012. "A lexically biased attentional set compensates for variable speech quality caused by pronunciation variation." *Language and Cognitive Processes* 27 (7-8): 1225-1239.

- Polka, Linda, and Janet F Werker. 1994. "Developmental changes in perception of nonnative vowel contrasts." *Journal of Experimental Psychology: Human Perception and Performance* 20 (2): 421.
- Polka, Linda, Connie Colantonio, and Megha Sundara. 2001. "A cross-language comparison of /d/-/ð/ perception: evidence for a new developmental pattern." *The Journal of the Acoustical Society of America* 109 (5): 2190-2201.
- Repp, Bruno H. 1981. "Two strategies in fricative discrimination." *Perception & Psychophysics* 30 (3): 217-227.
- Rost, Gwyneth C., and Bob McMurray. 2010. "Finding the Signal by Adding Noise: The Role of Noncontrastive Phonetic Variability in Early Word Learning." *Infancy* 15 (6): 608-635. Accessed 11 15, 2018. <http://onlinelibrary.wiley.com/doi/10.1111/j.1532-7078.2010.00033.x/full>.
- Rost, Gwyneth C., and Bob McMurray. 2009. "Speaker variability augments phonological processing in early word learning." *Developmental Science* 12 (2): 339-349. Accessed 11 15, 2018. <https://ncbi.nlm.nih.gov/pubmed/19143806>.
- Saffran, Jenny R, Elissa L Newport, Richard N Aslin, Rachel A Tunick, and Sandra Barrueco. 1997. "Incidental language learning: Listening (and learning) out of the corner of your ear." *Psychological Science* 8 (2): 101-105.
- Saffran, Jenny R, Richard N Aslin, and Elissa L Newport. 1996. "Statistical learning by 8-month-old infants." *Science* 274 (5294 ): 1926-1928.
- Schnoebelen, Tyler, and Victor Kuperman. 2010. "Using Amazon Mechanical Turk for linguistic research." *Psihologija* 43 (4): 441-464.
- Schwartz, Bonnie D, and Rex A Sprouse. 1996. "L2 cognitive states and the full transfer/full access model." *Second Language Research* 12 (1): 40-72.
- Scott, Lisa S, and Alexandra Monesson. 2009. "The origin of biases in face perception." *Psychological Science* 20 (6): 676-680.
- See, Judi E, Joel S Warm, William N Dember, and Steven R Howe. 1997. "Vigilance and signal detection theory: An empirical evaluation of five measure of response bias." *Human Factors* 39 (1): 14-29.
- Seidl, Amanda, Alejandrina Cristia, Amelie Bernard, and Kristine H Onishi. 2009. "Allophonic and phonemic contrasts in infants' learning of sound patterns." *Language Learning and Development* 5 (3): 191-202.
- Seidl, Amanda, and Alejandrina Cristia. 2012. "Infants' learning of phonological status." *Frontiers in Psychology* 3: 448.
- Shapiro, Danielle N, Jesse Chandler, and Pam A Mueller. 2013. "Using Mechanical Turk to study clinical populations." *Clinical Psychological Science* 1 (2): 213-220.
- Sharma, Anu, and Michael F Dorman. 1999. "Cortical auditory evoked potential correlates of categorical perception of voice-onset time." *The Journal of the Acoustical Society of America* 106 (2): 1078-1083.

- Sowden, Paul T, Ian R. L Davies, and Penny Roling. 2000. "Perceptual learning of the detection of features in X-ray images: A functional role for improvement in adults' visual sensitivity?" *Journal of Experimental Psychology: Human Perception and Performance* 26 (1): 379-390.
- Stager, Christine L, and Janet F Werker. 1997. "Infants listen for more phonetic detail in speech perception than in word-learning tasks." *Nature* 388 (6640 ): 381.
- Stanislaw, Harold, and Natasha Todorov. 1999. "Calculation of signal detection theory measures." *Behavior Research Methods, Instruments, & Computers* 31 (1): 137-149.
- Strand, Elizabeth A, and Keith Johnson. 1996. "Gradient and visual speaker normalization in the perception of fricatives." *KONVENS* 14-26.
- Stretch, Vincent, and John T Wixted. 1998. "Decision rules for recognition memory confidence judgments." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 24 (6): 1397-1410.
- Sundara, Megha, Linda Polka, and Fred Genesee. 2006. "Language-experience facilitates discrimination of /d-ð/ in monolingual and bilingual acquisition of English." *Cognition* 100 (2): 369-388.
- Swets, John A. 2014. *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. New York: Psychology Press.
- Swingley, Daniel. 2009. "Contributions of infant word learning to language development." *Philosophical Transactions of the Royal Society B: Biological Sciences* 364 (1536): 3617-3632.
- Swingley, Daniel. 2007. "Lexical exposure and word-form encoding in 1.5-year-olds." *Developmental Psychology* 43 (2): 454.
- Swingley, Daniel, and Richard N Aslin. 2000. "Spoken word recognition and lexical representation in very young children." *Cognition* 76 (2): 147-166.
- Tanner Jr, Wilson P, and John A Swets. 1954. "A decision-making theory of visual detection." *Psychological Review* 61 (6): 401.
- Team, R Core. 2014. "R: A language and environment for statistical computing."
- Ter Schure, Sophie, Caroline Junge, and Paul Boersma. 2016. "Discriminating non-native vowels on the basis of multimodal, auditory or visual information: Effects on infants' looking patterns and discrimination." *Frontiers in Psychology* 7: 525.
- Thiessen, Erik D. 2007. "The effect of distributional information on children's use of phonemic contrasts." *Journal of Memory and Language* 56 (1): 16-34.
- Thiessen, Erik D, and Jenny R Saffran. 2003. "When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants." *Developmental Psychology* 39 (4): 706.
- Toro, Juan M, Marina Nespor, Jacques Mehler, and Luca L Bonatti. 2008. "Finding words and rules in a speech stream: Functional differences between vowels and consonants." *Psychological Science* 19 (2): 137-144.

- Toro, Juan M, Scott Sinnett, and Salvador Soto-Faraco. 2005. "Speech segmentation by statistical learning depends on attention." *Cognition* 97 (2): B25-B34.
- Trehub, Sandra E. 1976. "The discrimination of foreign speech contrasts by infants and adults." *Child Development* 466-472.
- Tsao, Feng-Ming, Huei-Mei Liu, and Patricia K Kuhl. 2006. "Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants." *Journal of the Acoustical Society of America* 120 (4): 2285-2294.
- Vallabha, Gautam K, James L McClelland, Ferran Pons, Janet F Werker, and Shigeaki Amano. 2007. "Unsupervised learning of vowel categories from infant-directed speech." *Proceedings of the National Academy of Sciences* 104 (33): 13273-13278.
- Von Holzen, Katie, and Christina Bergmann. 2018. "A Meta-Analysis of Infants' Mispronunciation Sensitivity Development." *40th Annual Conference of the Cognitive Science Society*. 1159-1164.
- Vouloumanos, Athena, and Janet F Werker. 2004. "Tuned to the signal: the privileged status of speech for young infants." *Developmental Science* 7 (3): 270-276.
- Vouloumanos, Athena, and Janet F. Werker. 2007. "Listening to language at birth: Evidence for a bias for speech in neonates." *Developmental Science* 10 (2): 159-164.
- Wanrooij, Karin, Paola Escudero, and Maartje EJ Raijmakers. 2013. "What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning." *Journal of Phonetics* 41 (5): 307-319.
- Wanrooij, Karin, Paul Boersma, and Titia L van Zuijen. 2014. "Distributional vowel training is less effective for adults than for infants. A study using the mismatch response." *PLOS ONE* 9 (10): e109806.
- Weber, Andrea, and Anne Cutler. 2004. "Lexical competition in non-native spoken-word recognition." *Journal of Memory and Language* 50 (1): 1-25.
- Werker, Janet F, and Chris E Lalonde. 1988. "Cross-language speech perception: Initial capabilities and developmental change." *Developmental Psychology* 24 (5): 672.
- Werker, Janet F, and Richard C Tees. 1984. "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life." *Infant Behavior and Development* 7 (1): 49-63.
- Werker, Janet F, and Richard C Tees. 1983. "Developmental changes across childhood in the perception of non-native speech sounds." *Canadian Journal of Psychology* 37 (2): 278.
- Werker, Janet F, and Suzanne Curtin. 2005. "PRIMIR: A developmental framework of infant speech processing." *Language Learning and Development* 1 (2): 197-234.
- Werker, Janet F, Ferran Pons, Christiane Dietrich, Sachiyo Kajikawa, Laurel Fais, and Shigeaki Amano. 2007. "Infant-directed speech supports phonetic category learning in English and Japanese." *Cognition* 103 (1): 147-162.

- Werker, Janet F, H. Henny Yeung, and Katherine A Yoshida. 2012. "How do infants become experts at native-speech perception?" *Current Directions in Psychological Science* 21 (4): 221-226.
- Werker, Janet F, John HV Gilbert, Keith Humphrey, and Richard C Tees. 1981. "Developmental aspects of cross-language speech perception." *Child Development* 349-355.
- Werker, Janet F., Christopher T Fennell, Kathleen M Corcoran, and Christine L Stager. 2002. "Infants' ability to learn phonetically similar words: Effects of age and vocabulary size." *Infancy* 3 (1): 1-30.
- Winawer, Jonathan, Nathan Witthoft, Michael C Frank, Lisa Wu, Alex R Wade, and Lera boroditsky. 2007. "2007." 104 (19): 7780-7785.
- Xie, Xin, Rachel M Theodore, and Emily B Myers. 2017. "More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories." *Journal of Experimental Psychology: Human Perception and Performance* 4.
- Yoshida, Katherine A, Christopher T Fennell, Daniel Swingley, and Janet F Werker. 2009. "Fourteen-month-old infants learn similar-sounding words." *Developmental Science* 12 (3): 412-418.
- Yoshida, Katherine A, Ferran Pons, Jessica Maye, and Janet F Werker. 2010. "Distributional phonetic learning at 10 months of age." *Infancy* 15 (4): 420-433.
- Yuan, Jiahong, and Mark Liberman. 2011. "Automatic measurement and comparison of vowel nasalization across languages." *Proceedings of ICPHS*. Hong Kong. 2244-2247.
- Zhang, Hang. 2013. *The second language acquisition of Mandarin Chinese tones by English, Japanese and Korean speakers*. Doctoral dissertation, Chapel Hill, NC: The University of North Carolina at Chapel Hill.
- Zsiga, Elizabeth C. 2013. *The Sounds of Language: An Introduction to Phonetics and Phonology*. Chichester, West Sussex: John Wiley & Sons.