Perceptual adaptation to non-native speech:
The effects of bias, exposure, and input variation

by

Rebecca Laturnus

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Linguistics
New York University
May, 2018

Lisa Davidson
Dedication

For Eric Vatikiotis-Bateson
Mentor, confidant, critic, friend
You’ll always be my favourite uncle
Acknowledgements

Thank you to each of my participants, who patiently gave up an hour of their time, recruited their friends and total strangers, and made this project possible. Thank you to the community of linguists who recruited and recorded native speakers of regional English varieties.

Thank you to my committee members, whose incisive feedback, questions and advice were invaluable in bringing the project to fruition. I owe immense gratitude in particular to my chair, Lisa Davidson, whose inhumanly quick feedback, guidance, unending patience, and willingness to let me pursue the projects that interested me were transformative for my graduate school experience. Thanks to Greg Guy for his acute observations and insightful questions on every project. They have made me a better linguist. I have enjoyed all of our conversations over the years. Renee Blake pushed me to think more deeply about patterns and questions in my data that would never have occurred to me, and the project has benefitted immensely from our interactions. Melissa Baese-Berk wrote the paper that started it all. Her enthusiasm, willingness to share notes and offer suggestions inspired this dissertation from its infancy. Susie Levi offered methodological advice that improved the project greatly. Her perspective was invaluable throughout the process.

Before I was a confused grad student, I was an even more lost and confused college kid. The Department of Linguistics at UBC truly made my undergraduate experience. I'll always think of it as my home away from home. I am grateful for each of the professors with whom I interacted: Eric Vatikiotis-Bateson, Molly Babel, Martina Wiltschko, Lisa Matthewson, Hotze
Rullmann, Doug Pulleyblank, Gunnar Hansson, Henry Davis, Kathleen Currie Hall, Joseph Stemberger, Bryan Gick, Carla Hudson Kam, Michael Rochemont, and Rose-Marie Dechaine. Thank you for involving me in research and department life and for all of your encouragement. Molly’s instrumental phonetics class is where I first fell in love with the field. She's been a role model and inspiration and I've appreciated her advice and friendship over the years. Martina provided my first thorough introduction to documenting understudied languages and gave me opportunities to keep researching once classes were over. Lisa and Hotze guided me to a more thorough understanding of the topics I was interested in. I continue to appreciate their faith and expectation that I could keep up. Eric was Eric and I miss him every day.

I met some amazing people at NYU. Zack Jaggers and Ben Yancer were the best neighbours and friends I could ever ask for. Our dinners in Flatbush are some of my favourite memories in the city. Having Suzy Ahn as an office mate almost made me want to come in every day (almost). I've loved and valued our commiserations and conversations since we both moved away. Shannon Ward and Eli Dollarhide provided a necessary respite from linguistics and made transitioning to life in the city so much easier. I count myself lucky to call them friends. Zack, Adina Williams, Dan Duncan, Yohei Oseki, Isaac Bleaman, Songhee Kim, and Jeff Lin ensured first year was an enjoyable, collaborative experience. The food wasn't always great, but the company was stellar. They've continued to be great friends and confidants ever since. Ildi Szabo and Maddie Gilbert checked in on me every so often this past year. I appreciated their keeping tabs on me and I've enjoyed getting to know them better. Many other people have contributed to the sense of community I've felt in the department, including Mary Robinson, Dunja
Veselinovic, Yining Nie, Natalie Povilonis de Vilchez, Allison Shapp, Sheng-Fu Wang, Sarah Phillips, Nicole Holliday, Dani Szeredi, James Whang, and Sean Martin. I've enjoyed and appreciated their friendship. I want to thank in particular those who went before me: Dani and Nate LaFave for giving me hope in life after grad school, Sean and James for often listening to my methodological concerns and lending their expertise, and Nicole for unwittingly being my role model for years, making me a better person and linguist. You keep writing, I'll keep reading.

Moving across a different country was an adventure, and it's unlikely I'd have stuck it out without my people back home. I'm grateful to my parents and brothers, for keeping me grounded, for giving me perspective, and for helping me maintain my sense of humour even when I wanted to burn it all down. My sisters-in-law have offered years of much-needed emotional support and encouragement, to me and to the rest of the pack. We're so much better off with each of them around. Jude, Charlotte and Frankie are a constant source of hilarity and amazement. I'm so proud to be their Apple Becky. I'm thankful to Pablo Vargas for our nights of laughing so hard it hurts, for movie dates, and for always telling me I'm a badass, even when I feel useless. I'm thankful for Liv Yoon, whose friendship has made a world of difference this past year in particular. Our VPL writing dates made the process much less terrible. Thanks for being one of the best, most generous people I know. Janice D’Souza, Andrea Perez-Leon Anaya, and Lauren Morrow have put up with my endless stream-of-consciousness venting and running commentary nearly every day for over 10 years. They always know exactly what to say. Our friendship had me convinced I had a normal, healthy social life even when I didn't leave my
apartment for 7 days at a time (no regrets!). Thanks for always being there, guys. Andrea, thanks for suggesting we take that LING 101 class all those years ago.

I would never have made it to grad school without John-Jose Nuñez and Emily Blamire. Em, thanks for getting me through second year alive, and then deciding to thrive during third year when I was in Quebec. I might have never gotten my act together without you making me feel left behind. Thanks for making sure tri-annual visits happen and not letting me work the whole time. Thanks for introducing me to musicals. John-Jose, thanks for putting up with the roller coasters, doing all the cooking, and never being annoyed when I work(ed?) too much. Thanks for sneakily getting me out of that a French lit exam and not asking me first. Thanks to both of you for sharing custody, for monitoring my health, for always believing in me, and for encouraging me to be more assertive and try new things. Thanks for being awesome.

Special shoutout to all the people who have gone out of their way to feed me over the years: Emily, John-Jose, Zack and Ben, Liv, Mama and Papa Blamire, the Nuñez/Del Medico clan, and Ma & Pa since birth. Showing love through food is a language I understand well because I've been on the receiving end so very often. Love you all!
Abstract

Previous research suggests listeners’ expectations of the speech signal can affect their comprehension and perception of unfamiliar voices. The literature has also demonstrated, however, that people can improve their comprehension of even difficult speech through high-variability exposure. Through a series of three experiments, this dissertation contributes several key findings to the growing body of literature investigating adaptation and generalization to unfamiliar talkers. Consistent with past work (Baese-Berk et al., 2013), the first experiment found listeners who were trained on five unrelated non-native (L2) accented talkers had more accurate comprehension of both a novel talker of an accent they had been trained on, and a novel talker of an unfamiliar accent. These data provide further evidence that experience with a diversity of L2 accents facilitates generalization.

Experiment 2 tested whether this finding could be replicated with real-world exposure to accented speech, rather than the artificial laboratory training that is common to attunement research. Individuals with greater lifetime experience listening to non-native accented speech, particularly from family, friends, and in their childhood community, were found to have more accurate comprehension of two unfamiliar L2 talkers compared to listeners with less prior exposure to non-native speech.

The third experiment investigated whether exposure to multiple different accents of any kind would provide listeners with enough variability of experience to facilitate generalization to non-native voices. Training on five different native regional varieties of English did not facilitate
comprehension of unfamiliar non-native talkers, however, and no difference was found compared to a control group trained on five speakers of American English. This suggests it is familiarity with the intra-talker variation characteristic of non-native speech that aids in generalization to other L2 talkers, and not just experience with inter-talker variation that differs from the listener’s own dialect.

Finally, people with more negative biases towards non-native speech were less accurate in their transcriptions of unfamiliar L2 talkers. This pattern was observed for all participant groups in all three experiments, with the exception of individuals who had extensive prior lifetime experience interacting with non-native speakers. Though this group of listeners still exhibited the same range of biases as other conditions, biases were not found to be predictive of comprehension accuracy.

Together these results are consistent with a model of adaptation and generalization wherein listeners learn patterns common to L2 speech through exposure, including the structures that are often difficult for L2 speakers and the tendency for greater within-speaker variation. Greater knowledge of these patterns is proposed to facilitate comprehension of novel non-native talkers, rather than reliance on new speech being acoustically similar to prior accented input.
## Contents

List of Figures xi

List of Tables xiii

1 Introduction 1

1.1 Listeners’ perception of speech variability 1

1.2 Adaptation to difficult speech 8

1.3 Adaptation to non-native speech 10

1.4 The effects of bias, exposure, and input variation in accent attunement 30

2 Implicit bias and accent attunement 34

2.1 Methods 37

2.2 Results 50

2.3 Discussion 58

2.4 Conclusion 63

3 Greater lifetime exposure facilitates generalization 65

3.1 Methods 66

3.2 Results 71

3.3 Discussion 77

3.4 Conclusion 78

4 Intra- versus interspeaker variation 79
List of Figures

Figure 1. Comparison of methods for Experiment 1 and Baese-Berk et al. (2013) 41
Figure 2. Examples of IAT task screens 44
Figure 3. Proportion accuracy on test trials by training group in Experiment 1 53
Figure 4. Number of errors by training block 54
Figure 5. Plot of accuracy on test trials by IAT score for Experiment 1 55
Figure 6. Scatter plot of IAT scores by accuracy for Experiment 1 56
Figure 7. Proportion accuracy on orthographic IAT trials for all participants 57
Figure 8. Proportion accuracy on test trials by condition in Experiment 2 73
Figure 9. Plot of accuracy on test trials by IAT score in Experiment 2 75
Figure 10. Scatter plot of IAT scores by accuracy for Experiment 2 76
Figure 11. Average VOT duration by English variety 87
Figure 12. Vowel plot of Chinese-accented English and American English. 89
Figure 13. Vowel plot of Korean-accented English and American English. 90
Figure 13. Vowel plot of Thai-accented English and American English. 90
Figure 15. Vowel plot of Russian-accented English and American English. 91
Figure 16. Vowel plot of Farsi-accented English and American English 91
Figure 17. Vowel plot of Italian-accented English and American English. 92
Figure 18. Vowel plot of Indian English and American English 93
Figure 19. Vowel plot of Jamaican English and American English 94
Figure 20. Vowel plot of Manchester English and American English 95
Figure 21. Vowel plot of New Zealand English and American English 96
Figure 22. Vowel plot of Scottish English and American English 97
Figure 23. Plot of unstressed vowels for native English varieties 99
Figure 24. Proportion accuracy on test trials by training group in Experiment 3 104
Figure 25. Number of errors by block in Experiment 3 105
Figure 26. Plot of accuracy on test trials by IAT score in Experiment 3 107
Figure 27. Scatterplot of IAT by proportion accuracy in Experiment 3 108
List of Tables

Table 1. List of talkers for acoustic analyses 22
Table 2. List of model talkers used during Experiment 1 39
Table 3. Results of pre-screening for semantic neutrality for the IAT 46
Table 4. Summary of answers on the Social Network Questionnaire 52
Table 5. Number of tokens extracted by place of articulation and voicing 82
Table 6. Total number of vowel tokens extracted per vowel for each speaker 83
Table 7. Average VOT duration and standard deviation for voiced tokens 86
Table 8. Average VOT duration and standard deviation for voiceless tokens 86
Table 9. Average schwa duration and standard deviation for native and L2 talkers 98
Table 10. Responses to Social Network Questionnaire for all three experiments 103
Chapter 1: Introduction

1.1 Listeners’ perception of speech variability

Humans are able to communicate and understand one other despite the incredible amount of variation present in the speech stream. The acoustic signal heard in conversation includes numerous talker-specific characteristics of the interlocutors. These include sex, gender, age, dialect, language background, and vocal tract physiology, as well as individuals’ own styles, identities, and idiolectal characteristics (e.g. Chambers et al., 2002; Drager, 2010; Eckert & Rickford, 2001; Holliday, 2016; Irvine, 2001; Johnson & Mullennix, 1997; Labov, 1972). Despite this inescapable variability, communication proceeds efficiently, even between strangers. Research in speech recognition suggests listeners retain information about the variability they’ve heard in memory, which facilitates understanding of similar voices.

The benefit of familiarity is well-documented in speech perception. Many studies have found people are sensitive to fine-grained phonetic details of specific talkers they have experience with in an experimental setting and seem to retain this information in memory (e.g. Bradlow et al., 1999; Church & Schacter, 1994; Cole et al., 1974; Craik & Kirsner, 1974; Geiselman & Bellezza, 1976; Goldinger, 1996; Goldinger et al., 1991; Levi et al., 2011; Nygaard & Pisoni, 1998; Nygaard & Queen, 2000; Nygaard et al., 1995; Schacter & Church, 1992; Walker et al., 1995). Palmeri, Goldinger, and Pisoni (1993) found people were faster and more accurate at identifying words as repeated in a continuous recognition memory experiment when the words were repeated in the same voice as they were initially presented; recognition was slower and less accurate when words were repeated in a different voice, regardless of whether the gender of the repeating voice matched
the initial presentation. Similarly, Bradlow, Nygaard, and Pisoni (1999) found listeners were more accurate recognizing previously-heard words when they were presented in the same voice and in the same speaking rate as they had originally heard them. Eisner and McQueen (2005) conducted a series of perceptual learning experiments, in which listeners hear a model talker produce an ambiguous token embedded in words that lexically bias the listener towards one interpretation or the other. In these experiments, Dutch participants heard fricatives ambiguous between [f] and [s] and were in a condition that either biased them to interpret the fricative as [f] via [f]-final words like *olijf* (‘olive’), or to interpret the fricative as [s] (e.g. *radijs*, ‘radish’). They found that participants in the [f]-biased condition categorized more sounds after exposure as [f] on an [ɛf- ɛs] continuum than did [s]-biased participants. This pattern emerged when the continuum used the exposure talker’s speech, but not when it was based on another model talker.

Listeners’ expectations for the speech signal have also been shown to affect speech perception (e.g. age: Drager, 2011; Hay et al., 2006; Koops et al., 2008; gender: Johnson et al., 1999; Strand & Johnson, 1996; style: Eckert & Rickford, 2001; Irvine, 2001; socioeconomic background: Hay et al., 2006; race: Casasanto, 2008). Walker and Hay (2011) found age effects in an auditory lexical decision task by using vocabulary that differed in their frequency of use among different generations. When listeners heard a word they associated with an older generation (e.g. *sixpence, petticoat*) produced by an older speaker, they were faster and more accurate in recognizing it. This was also true for young words (e.g. *Internet, sexist*) heard in a younger voice. When the input was congruent with their expectations, therefore, participants were faster and more accurate in processing it. Relatedly, Hay et al. (2006) investigated how the social factors of perceived age and class affected perception of a merger-in-progress in New Zealand. At the time
of writing, the merger was most strongly associated with younger and working-class individuals. Participants saw a picture before each audio stimulus and were instructed the picture was of the person saying the word they were about to hear. The pictures of individuals differed in their apparent age or social class, being either older or younger, middle class or working class. They found participants who did not have the vowel merger under investigation had different perceptions of the stimuli depending on the age they believed the talker to be: after seeing a photo of an older speaker, participants were more accurate categorizing words, while seeing a photo of a younger speaker caused them to treat words as more ambiguous, reducing accuracy rates. For the class-distinct photos, participants were most accurate for the voice whose vowels were the most distinct paired with the highest-class photo. All stimuli were recorded by individuals who reliably made a distinction between the two vowels undergoing merger, suggesting the photos did affect perception for those who themselves pronounced the two vowels differently.

In another study manipulating audiovisual cues, Strand (1999) found listeners’ perception of /s/ and /ʃ/ were affected by the gender of the speaker in both an audio-only experiment and in an audiovisual experiment. In the audio experiment, four voices (two male, two female) were selected who had been rated as the most or least prototypical of their sex from among a group of 37 voices, such that the four voices chosen were one prototypical male, one non-prototypical male, one prototypical female, and one non-prototypical female. Fricatives were synthesized for each model talker on a continuum from /ʃ/ to /s/ and combined with natural vowel-consonant sequences for each talker from the words sod and shod. Participants were then asked to identify each word they heard as either sod or shod. Results indicate that listeners make use of gradient expectations within the sex categories when perceiving these fricatives: the more prototypical the speaker, the
stronger the normalization of the fricative to conform with typical distributions for /ʃ/ and /s/ for native speakers of American English. A similar effect was found when the audio was accompanied by a picture of a male or female face: seeing the male face caused listeners’ boundary between /ʃ/ and /s/ to lower in frequency, while seeing the female face had the opposite effect.

Apparent dialect background suggested by visual cues can additionally influence how speech is perceived. Hay and Drager (2010) found participants’ perceptions of the speech they heard were influenced by exposure to a concept associated with a dialect area. Prior to beginning the experiment, participants were exposed to a stuffed animal that was either a kangaroo or koala (associated with Australia) or a kiwi bird (associated with New Zealand). Exposure happened by the experimenter affecting surprise at finding the stuffed animal in a cabinet with the response forms and moving the toy to the desk where the participants sat to complete the experiment. Participants were then asked to match vowels recorded by a male native speaker of New Zealand English to vowels along a synthesized continuum. The range of the continuum was more Australian-sounding (raised and fronted /I/) on one end and more New Zealand-sounding (lowered and centralized /I/) on the other. Participants perceived the vowels they heard as sounding more like whichever country’s dialect they had been primed to hear via seeing the stuffed animals. Niedzielski (1999) found a similar effect with listeners from Detroit. By writing either “Canada” or “Detroit” at the top of participants’ response forms, experimenters biased participants towards perceiving speech as sounding more like the stereotype of one location or the other. Listeners were asked to match different synthesized vowels with the model talker they heard, depending on whether they had been led to believe the talker was from Detroit or Canada. Results demonstrated their responses were based on how they believed people from those places sounded, more strongly
associating Canadian Raising with Canadians, even though it is a feature of Detroiter’s speech as well. Studies like these ones show listeners’ expectations about their interlocutors’ speech can bias how they perceive the acoustic signal.

Several studies have demonstrated that being led to believe a talker is of Asian descent can affect listeners’ perception in audiovisual tasks. In one experiment, Rubin (1992) presented a static picture of a Chinese woman or of a white woman alongside the voice of a native English speaker and found participants rated the voice as more accented and had worse accuracy when they saw the picture of the Chinese woman (see also Fraser & Kelly, 2012). In a similar study, Kang and Rubin (2009) report that participants who rated non-native speakers as less socially attractive (i.e. unfriendly, cold, hostile, dishonest), were less accurate in their comprehension when they thought the model talker was Asian, due to audiovisual stimuli, even though the model talker was again a native English speaker. Rather than finding correlations between bias or explicit attitudes and processing costs, Babel and Russell (2015) reported participants with more Asian Canadians in their personal social network showed a greater cost in comprehension. Their listeners rated native speakers of Canadian English (who were either Chinese Canadian or White Canadian) for accentedness on a 1 to 9 scale. In the audiovisual condition a picture of the talker was presented alongside the audio. They found listeners rated the Chinese Canadian voices as more accented in the audiovisual condition than in audio-only, despite these talkers being native speakers of Canadian English. The authors suggested a stronger expectation on the part of the listeners for the Chinese faces to have an accent may have biased their perception, leading them to believe these voices were more accented. Similarly, Yi, Phelps, Smiljanic, and Chandrasekaran (2013) found those with stronger Asian-foreign biases, as measured by an Implicit Association Test (Greenwald,
McGhee, & Schwartz, 1998), rated Korean-accented voices as more accented in an audiovisual condition than in an audio-only condition. Yi, Smiljanic, & Chandrasekaran (2014) investigated how listener bias to these same Asian faces affected speech perception using fMRI. They found that seeing the face of the Korean-accented talker exaggerated how accented participants perceived the voice to be, and also that greater Asian-foreign bias was associated with decreased neural efficiency in the audiovisual condition for the accented voices, suggesting additional processing resources were needed. Together, these studies suggest those with stronger expectations for individuals of Asian descent to have a strong non-native accent affects how they perceive these individuals’ speech.

1.1.1 Variation in non-native (L2) speech

Foreign accents are the result of certain sound patterns from an individual’s first language influencing the pronunciation of their second language (e.g. Archibald, 1998; Best, 1995; Blair & Ingram, 2003; Flege, 1995; Flege et al., 1995; Iverson et al., 2003; Kuhl, 1993; LaCharité & Paradis, 2005; Nguyen et al., 2008). Because of these influences from other languages, and because non-native speakers are less consistent in their pronunciation (Flege et al., 1995; Laturnus, Submitted; Rogers et al., 2012), it can be difficult for listeners to understand non-native speech, causing a loss of accuracy in comprehension (e.g. Donselaar & Lentz, 1994; Ferguson et al., 2010; Flege, 1984; Gass & Varonis, 1984; Munro & Derwing, 1995) and slower word recognition (e.g. Floccia et al., 2009; Munro & Derwing, 1995; Schmid & Yeni-Komshian, 1999; Weil, 2003; Wilson & Spaulding, 2010). Research suggests non-native speakers’ age of acquisition of their second language (e.g. Birdsong, 2005; Birdsong & Molis, 2001; Flege et al., 1995; Johnson & Newport, 1989; Marinova-Todd et al., 2003), and their length of residency in a country where the
language is spoken (e.g. Asher & García, 1969; Flege & Fletcher, 1992; Flege et al., 1995; Flege et al., 1999; Purcell & Suter, 1980), greatly affect the level of accentedness perceived by native listeners. Nygaard et al. (2006) found longer vowel durations among non-native speakers compared to native speakers contributed to intelligibility and accentedness ratings, while Tzeng et al. (2016) found greater vowel dispersion in productions of high frequency, low neighbourhood density words were rated as less intelligible and more accented. Bent et al. (2007) found L2 speakers’ vowel productions that differed from native speakers’ affected listeners’ comprehension more than production errors made on consonants. They also found errors made at the beginnings of words had the largest cost. Zielinski (2008) monitored listeners’ transcription of Korean-, Mandarin-, and Vietnamese-accented English for parts of the signal they found particularly difficult to understand. She found atypical syllable stress pattern, non-standard segments in strong syllables, and the interaction of these two production errors in particular had the most detrimental effect on comprehension. Non-native speakers are also reported to be more variable in their speaking rate compared to native speakers (Baese-Berk & Morrill, 2015), and show greater between-talker variability in word durations but less within-talker variability, leading to greater accentedness ratings by native listeners (Baker et al., 2011).

Non-native speech therefore includes variability in the form of production errors and transference effects from other languages, in addition to the variation from gender, sex, age, dialect background, and identity that is present in native speech. This makes the task of mapping from the speech stream to the intended lexical items even more difficult for listeners.

1.2 Adaptation to difficult speech

Despite the challenge presented by increased variability, research shows people are able to
improve their comprehension of difficult speech via exposure (Bent, Buchwald, & Pisoni, 2009; Borrie, McAuliffe, & Liss, 2012; Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Erb, Henry, Eisner, & Obleser, 2013; Francis, Nusbaum, & Fenn, 2007; Greenspan, Nusbaum, & Pisoni, 1988; Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008; Pallier, Sebastian-Gallés, Dupoux, Christophe, & Mehler, 1998; Peelle & Wingfield, 2005; Rosen, Faulkner, & Wilkinson, 1999; Schwab, Nusbaum, & Pisoni, 1985). Sheffert, Pisoni, Fellowes, and Remez (2002), for example, trained listeners on natural, sinewave, and reversed speech over the course of several days. Their findings include listeners learning to recognize individual talkers in all three speech types. They also observed that participants were able to generalize after training on sinewave speech to novel utterances, both in natural speech and in sinewave speech, demonstrating that even with unnatural speech types (e.g. reversed), listeners are able to adapt and generalize.

Bent et al. (2009) tested listeners’ ability to adapt to noise-vocoded speech, which simulates speech heard by cochlear implant patients, as well as to speech in multi-talker babble. They found that participants were better able to understand the noise-vocoded speech after training, and that there was consistency for which voices were the most intelligible in each condition. Listeners’ accuracy rates improved for both types of speech after a few minutes of exposure.

In the realm of accented speech, White and Aslin (2011) exposed 18- to 20-month-olds to artificial accents where vowels had been shifted in familiar words. They found that at test, toddlers were able to understand new words in the accent if they contained vowels that were shifted in the same way as those they had heard before (i.e. they were able to generalize to new items, but only if the items were consistent with the perceptual learning that happened during training). These
results are consistent with earlier work with adults conducted by Maye, Aslin, and Tanenhaus (2008), where participants heard a story in which front vowels were systematically shifted. During a subsequent lexical decision task, they showed greater acceptance of both familiar and unfamiliar words that were consistent with the shift. Like the toddlers in White and Aslin’s (2011) experiment, however, participants did not show greater acceptance of words containing vowel shifts that were inconsistent with what they heard during training.

Working with naturally accented speech, Clopper and Bradlow (2008) investigated the effect of background noise on listeners’ ability to understand and classify regional dialects. They found participants were able to adapt to variation when there was a moderate amount of noise, however higher amounts of background noise were detrimental to listeners of all dialect backgrounds. Adank, Evans, Stuart-Smith, and Scott (2009) investigated processing costs listeners face when listening to unfamiliar accents in noise. In an experiment testing participants from Southern England (London), they found listeners performed better on their own dialect in moderate noise than on a dialect they were less familiar with (Glaswegian English). The Glaswegian English speakers, in contrast, performed equally well on both dialects, which the authors suggest is due to these listeners having experience with both dialects, as Southern Standard British English (SE) is considered the ‘standard’ variety in the UK. A second experiment compared Londoners’ performance on Spanish-accented English in addition to their own dialect (SE) and Glaswegian English, all with various levels of background noise. They found listeners performed worst on the non-native Spanish-accented talkers, with lower accuracy and slower response times. The participants also had slower response times for Glaswegian English than for their own dialect at moderate noise levels.
Toddlers and children, like adults, have more trouble understanding unfamiliar accents than familiar ones (e.g. Barker & Turner, 2015; Bent, 2014; Bent & Atagi, 2017; Best et al., 2009; Mulak et al., 2013; Nathan et al., 1998; Schmale et al., 2011; Schmale & Seidl, 2009), but get better with exposure (Jones, Yan, Wagner, & Clopper, 2017; Potter & Saffran, 2017; Schmale, Cristia, & Seidl, 2012; Schmale, Cristia, Seidl, & Johnson, 2010; van Heugten & Johnson, 2014; van Heugten, Krieger, & Johnson, 2015; White & Aslin, 2011). Schmale et al. (2012) found 24-month-olds needed less than two minutes of exposure to adapt to an unfamiliar Spanish-accented talker. By 18 months, exposure to multiple regional, native accents (e.g. Scottish English, New Zealand English) facilitates comprehension of a novel regional accent (Potter & Saffran, 2017).

This advantage for exposure to multiple different accents has been found with adults as well. Clopper and Pisoni (2004) tested listeners’ ability to categorize unfamiliar talkers from each of six American English dialect regions following training. One group was trained on one model talker from each region, while a second group was trained on three talkers from each region. The group exposed to single talkers performed better during training and at test when they heard familiar voices, however the group exposed to three different voices from each area outperformed them when categorizing unfamiliar voices. These results suggest that greater inter-talker variability facilitates generalization to novel talkers.

1.3 Adaptation to non-native speech

This question of exposure to greater variability facilitating generalization is central to the current dissertation, specifically with respect to non-native speech. The present section provides an overview of the literature on adaptation and generalization to L2 speech, setting the stage for this projects’ specific research questions.
L2 accented speech presents a unique challenge for perception and processing. It contains variability not just due to the gender, age, dialect, style, and identity of the talker, but also variability stemming from being a late learner of the language. As reported for native and sinewave speech, however, listeners are able to improve their comprehension of an unfamiliar accent through exposure, usually following a brief period of adjustment. This adjustment, or adaptation, is known as attunement (Bradlow & Pisoni, 1999), when one goes from incomprehension of an accent to functional understanding.

Clarke and Garrett (2004) found listeners were able to attune to non-native accents rapidly, within two to four sentences or approximately one minute. Gass and Varonis (1984) investigated how different types of familiarity affected listeners’ comprehension of L2 speech. They found participants performed better when they were familiar with the topic of the utterances they were hearing. This was measured via accuracy on sentences for which a context had been supplied versus sentences with no supplied context. Their results also suggest that hearing a passage read by one non-native speaker facilitates 1) comprehension of that same speaker in novel utterances, 2) comprehension of another speaker of that accent, and 3) comprehension of another speaker of a different non-native accent. Levi (2015) trained 7- to 12-year-olds on the voices of German-English bilinguals over the course of five days. The children did show improvement for voices they were trained on, consistent with past research for adult listeners, but only for very familiar (i.e. high frequency) words. Despite training on three different German-accented voices, children did not show generalization to novel German-accented talkers at test. This could be due either to the nature of the stimuli, being single words in isolation rather than full sentences, or it could be due to the task during training, which required listeners to identify familiar voices, rather than
focusing attention on the content of what the voices were saying. Children did receive feedback during training on whether they had identified the voices correctly, but there was no requirement that they identify the words themselves, which may be required for talker-independent learning of accented speech.

Weil (2001) found the kind of exposure participants received affected their ability to generalize to new talkers as well. Specifically, he found generalization occurred between two talkers with the same first language only when listeners heard full sentences from model talkers. The suggested explanation is that sentences enable listeners to map variable input to lexical items, thereby learning the intended categories. This information is not available with single words unless listeners are provided with feedback on their interpretations of the signal. When Weil’s participants were trained and tested on single words in isolation, no feedback was provided, such that they had no indication of whether they had correctly interpreted the accented speech. In contrast, Sidaras, Alexander, and Nygaard (2008) trained participants on single-word Korean-accented stimuli but did provide them with feedback on their transcriptions, and listeners performed better on both familiar and unfamiliar Korean-accented talkers at test than did the untrained control groups.

When listeners do hear full sentences produced by accented talkers, research on perceptual adaptation to non-native speech suggests that high-variability training does greatly facilitate learning generalization. This is in line with what Clopper and Pisoni (2004) demonstrated for native regional accents. In the realm of second language acquisition, training listeners on multiple speakers has been shown to facilitate learners acquiring and retaining segmental contrasts (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Lively, Logan, & Pisoni, 1993; Wang, Jongman, & Sereno, 2003; Wang, Spence, Jongman, & Sereno, 1999).
Bradlow and Bent (2008) used sentential stimuli during their training paradigm and found listeners who were trained on a single foreign-accented voice did not generalize their learning to a novel talker with the same foreign accent at test, however those trained on multiple Chinese-accented speakers were able to generalize their learning to a novel speaker of Chinese-accented English. Bradlow and Bent additionally included a test talker in this experiment who was a native speaker of Slovakian, but they found training on the five Chinese-accented speakers did not generalize to the Slovakian speaker at test. As a follow up to this research, Sidaras, Alexander, and Nygaard (2009) conducted several experiments investigating perceptual adaptation and generalization to non-native talkers. Participants in the first experiment were trained on 3 male and 3 female Spanish-accented talkers producing simple sentences. At test, listeners heard either the same set of talkers or a different set of 6 non-native Spanish voices, with the expectation that if listeners learned in a talker-specific way, they would only have improved performance on familiar talkers, while learning in a talker-independent way should generalize to the 6 novel talkers. They found listeners generalized to both novel sentences and to novel talkers. Experiment 2 trained and tested participants on words instead of sentences. The authors predicted that improved performance at test would indicate listeners were able to learn segmental information, since the suprasegmental information available in sentences would not be available for single word stimuli. Crucially, feedback during training was provided. They found learning did generalize to both novel words and novel talkers. A close analysis of participants’ transcriptions of difficult vowels (i.e. vowels not present in Spanish’s inventory) revealed those who were trained on Spanish-accented English improved their accuracy in recognizing a subset of the highly confusable vowels with training, compared to participants who did not get training. They additionally found those vowels
that did improve were more acoustically distinct in the productions of the non-native talkers from the vowels that did not improve. The authors argue this suggests listeners were able to use these acoustic cues to learn patterns of systematic variation present in the Spanish accent and generalize their knowledge to novel talkers at test, even in single word utterances with no feedback provided.

1.3.1 Generalizing across accents: Baese-Berk et al. (2013)

Baese-Berk, Bradlow and Wright (2013) further tested the efficacy of high-variability training by exposing listeners to five different speakers of five unrelated accents. Participants were then tested on a new speaker of an accent heard in training and a new speaker of an untrained accent. During training, participants heard short sentences embedded in background noise at a signal-to-noise ratio of +5dB. The sentences were recordings of two lists of Bamford-Kowal-Bench sentences (Bench, Kowal, & Bamford, 1979) from the Northwestern University Foreign-Accented English Speech Database (NUFAESD) (Bent & Bradlow, 2003). Each list was used in a separate training session, which happened on consecutive days. During each session, participants heard five repetitions of each sentence in the list, with each repetition produced by a different non-native accented talker (Hindi, Korean, Mandarin, Russian, Thai). The model talkers, all male, were rated for intelligibility by 40 native speakers listening to BKB sentences read by the talker embedded in background noise. Intelligibility was calculated as the percentage of keywords participants correctly transcribed for that speaker. The model talkers chosen for Baese-Berk et al.’s experiment had mid-range intelligibility scores.

During each of two test blocks, participants heard a new BKB list read by an unfamiliar talker embedded in background noise, which they had to transcribe. One test talker was a native
speaker of Mandarin (a different speaker from the one heard in training) and the other was a native speaker of Slovakian. Both test talkers were also males with mid-range intelligibility.

Accuracy on test blocks was measured as the percentage of keywords correctly transcribed. Participants’ performance was compared to the training groups from Bradlow and Bent’s (2008) study, where one condition was trained on five Chinese\(^1\)-accented talkers and the other was trained on five speakers of American English. The training materials in Bradlow and Bent’s experiment were identical to Baese-Berk et al.’s study, and the test blocks (sentences and model talkers) were also identical, allowing for comparisons across the two studies.

Baese-Berk et al. found adaptation was facilitated at test for both the novel Chinese-accented talker and for the Slovakian-accented talker following training on the five non-native voices. They significantly outperformed the Bradlow and Bent (2008) participant groups in transcribing the Slovakian-accented talker, whose training had been on five Chinese-accented voices. On the Chinese-accented test talker, they again outperformed the group trained on American English, and did not significantly differ in performance from the participants trained only on Chinese-accented talker. These findings suggest listeners who have greater experience with several speakers of multiple unrelated L2 accents may be better able to comprehend non-native speech in general.

### 1.3.2 How do listeners benefit from exposure to non-native accents?

Several key hypotheses have been proposed in the accent attunement literature to account

\(^1\) Bradlow and Bent (2008) refer to their model talkers during training and test as Chinese-accented, while Baese-Berk et al. (2013) refer to theirs as Mandarin-accented. Since comparisons were made between studies, the assumption here is that speakers in both studies were Mandarin-accented.
for how speakers adapt to non-native speech and generalize their learning to novel non-native talkers. One such explanation, the **Systematic Variability Hypothesis**, suggests there are commonalities in the productions of non-native speakers, regardless of their language background. The speech of non-natives tends to be slower and more halted than native speakers (Baker et al., 2011; Guion, Flege, Liu, & Yeni-Komshian, 2000; Munro & Derwing, 1995), and many of English's contrasts are difficult for speakers of other languages in similar ways. Patterns in speakers’ first languages that commonly affect English production include lacking certain English vowel contrasts (e.g. Flege, Yeni-Komshian, & Liu, 1999; Fox, Flege, & Munro, 1995; Oh et al., 2011; Tsukada, 2009; Wade, Jongman, & Sereno, 2007); lack of aspirated stops in their L1 (Nagy & Kochetov, 2013; Ringen & Kulikov, 2012); phonating during stop closures (Bortolini, Zmarich, Fior, & Bonifacio, 1995; Ringen & Kulikov, 2012; Shimizu, 2011); lacking interdental fricatives (Gonet & Pietron, 2006; Hanulikova & Weber, 2010; Rau, Chang, & Tarone, 2009); and having difficulty with English liquids (Bradlow, 2008; Bradlow et al., 1997; Flege, Takagi, & Mann, 1995, 1996; Lively et al., 1993; Riney & Flege, 1998) or English consonant clusters (Broselow & Finer, 1991; Eckman & Iverson, 1993; Kabak & Idsardi, 2007).

The difficulty of these structures means L2 speakers with different L1s might make similar production errors in English. When listeners are exposed to multiple different accents, this hypothesis suggests they’re able to generalize their attunement to a novel accent because the ways in which it differs from English are presumably similar to the ways other accents differ (Baese-Berk et al. 2013). There are (at minimum) two possible interpretations of this hypothesis, as it has been presented in the literature. The first view, a strong interpretation, would predict a high degree of acoustic overlap between individual speakers of different language backgrounds (i.e. different
speakers would find the same contrasts difficult, and resolve them in the same way acoustically). Using perceptual learning, listeners would closely track the distribution of acoustic cues for each of a talker’s categories. Generalization between different speakers would occur because listeners are able to draw on past experience with another accented speaker who is a close acoustic match for the listener’s current interlocutor.

Evidence for this view has been found in attunement studies using single word stimuli in Mandarin-accented English (Xie, 2015; Xie, Theodore, & Myers, 2017). Word-final coronal stops in this accent are pronounced as voiceless and reliably differ in burst length, but tend not to differ in the length of the preceding vowel. This makes them perceptually ambiguous to English listeners, who don’t usually make use of burst length as an acoustic category cue. In a lexical decision task, participants in these studies heard single-word stimuli ending in /t/ or /d/ produced by one speaker during training and a different speaker at test. The results show that generalization (via priming) occurred only when participants were trained on a voice whose /t/ and /d/ productions acoustically matched those of the test talker, suggesting bottom-up acoustic similarity was necessary for generalization. Xie and colleagues additionally found listeners reorganized their internal category structure for isolated /t/- and /d/-final words in response to hearing Mandarin-accented English (Xie et al., 2017), such that they were able to make category goodness judgements about speech tokens. Participants in Xie et al.’s (2017) experiment first heard Mandarin-accented, multisyllabic single words in isolation during an auditory lexical decision task. Among the stimuli were 30 /d/-final words. Following this exposure phase, they completed a practice phase where they became accustomed with making category goodness judgements using a different contrast of interest than word-final /d/ (word-final /m/ vs. /n/). For each word participants heard, they had to rate the
goodness of final consonant on a scale from “1 (very poor example of /m/) to 7 (very good example of /m/)” (Xie et al. 2017:212). The test phase used the same procedure, except that listeners heard /d/- and /t/-final words. Ratings were made on each item as a token of the category /d/ and of the category /t/ (ratings for each category were elicited in different blocks). Results revealed greater ratings for both /d/- and /t/-final words as exemplars of the appropriate categories, relative to the control group, who received no exposure to /d/-final words during training. The authors argue that because listeners’ ratings were affected for both the /t/ and /d/ categories when they only received training on /d/-final words, this should be taken as evidence that listeners adjusted the internal structure of their /t/ and /d/ categories by tuning into specific acoustic cues of the contrast (i.e. burst length), rather than simply shifting the /d/ category boundary to encapsulate heard tokens. Had listeners only shifted their /d/ category boundary, they argue, greater ratings for accented /t/ tokens as /t/ exemplars would not be expected.

Talker productions of the stimuli in Xie (2015) and Xie et al. (2017) were highly consistent, however. Other segments in L2 speech, such as vowels, can be much more variable (e.g. Laturnus, Submitted; Wade et al., 2007). This increased variability might make the task of tracking the distribution of acoustic cues for each category, and restructuring their own categories accordingly, a more difficult task for listeners. Furthermore, these studies exposed listeners to only single-word stimuli. Attunement tasks more often make use of sentential stimuli, because they have been shown to have a greater facilitating effect on generalization to novel speakers (e.g. Clarke & Garrett, 2004; Wade et al., 2007; Weil, 2001). Full sentences contain many more possible segments that could acoustically differ between voices.
An alternative to the strong view is the **Weak Systematic Variability Hypothesis**, which suggests that instead of close acoustic matches, perhaps greater experience with L2 speech gives listeners more subconscious knowledge of how non-native speech can deviate from native productions. Many non-native speakers, for example, lack the tense-lax contrast in English vowels, because it is not a feature of their native language. While two speakers of different accents may not implement their /e/ and /ɛ/ categories in acoustically similar ways, the lack of this contrast in one voice might speed word recognition of another talker who also lacks the contrast, even if there is no acoustic overlap in their /e/ or /ɛ/ productions. Increased awareness of which structures in English are difficult for L2 speakers (e.g. tense/lax vowel contrasts, interdental fricatives, or /ɹ/), or patterns common to L2 speakers (e.g. greater within-talker variability), might facilitate more rapid adaptation to novel L2 talkers who also have difficulty on those structures, without requiring the individual talkers to make acoustically similar production errors. This is in comparison to lacking knowledge of which structures are hard and having to begin adaptation from scratch. Of note with this hypothesis is that if listeners encounter a talker who deviates from English in ways that are *not* typical for other L2 talkers the listener has heard before, they will have a harder time with comprehension and will take longer to adapt to this new voice.

The Systematic Variability Hypothesis, in either its strong or weak form, suggests that listeners need exposure to non-native speech specifically, in order to generalize to novel non-native talkers. This is because in its strong form, listeners reorganize internal category structure to match that of the input, and only another non-native speaker is likely to be a close acoustic match for those categories. The proposal for category reorganization is that listeners track the acoustic input of each talker they hear and form new categories for each talker in a talker-specific model (Xie et
Hearing a talker who is a strong acoustic match for a previously-encountered talker therefore has a facilitating effect on comprehension because the relevant categories have already been adjusted to match the input. In the weak interpretation of the Systematic Variability Hypothesis, exposure to L2 speech allows listeners to learn which structures or patterns L2 speakers typically produce with maximum deviation from native English so that they are faster to adapt to new accented talkers. Other hypotheses that might account for generalization do not have this requirement of past experience with L2 speech.

An alternative to the two forms of the Systematic Variability Hypothesis is the **Category Loosening Hypothesis**, which suggests that listeners expand their category boundaries to encapsulate incoming variation without necessarily reorganizing internal category structure. Many studies have shown that listeners expand categories to encapsulate the variation they’ve been exposed to via ambiguous tokens embedded in words (e.g. Clarke-Davidson et al., 2008; Eisner & McQueen, 2006; Kraljic & Samuel, 2005, 2006; Maye et al., 2008; Schreiber et al., 2013; Sumner, 2011). If listeners store these boundary shifts in memory, new accent variation could fall within the already loosened category boundaries without constituting a close acoustic match to previously heard speech, leading to generalization. Crucially, boundary shifts under this hypothesis do not constitute a general relaxation of a category, but instead occur based solely on past bottom-up acoustic input, encapsulating only the variation a listener has heard before. Unlike the strong interpretation of the Systematic Variability Hypothesis, however, incoming speech can match to categories resulting from the aggregate of past experience, rather than matching one speaker closely. The Weak Systematic Variability Hypothesis, in contrast, tracks more general patterns
(e.g. tense/lax vowels are hard for L2 speakers, but the way they implement these categories often differ from each other), so acoustic similarity to past input isn’t strictly necessary.

1.3.3 Searching for systematic variability in non-native speech

To investigate the viability of these explanations of generalization, (Laturnus, Submitted) conducted a systematic acoustic comparison between three native speakers of American English and six non-native accents. The model talkers were recordings taken from Northwestern University’s Wildcat Corpus of Native- and Foreign-Accented Speech. This corpus was used as a close approximation of the database used by Baese-Berk et al. (2013), which, due to ethics restrictions, was not available for analysis. The Wildcat Corpus was recorded in the same location with comparable methods. It additionally recruited participants from a comparable population, the majority being newly-admitted graduate students to Northwestern University attending an English language and acculturation program during the summer prior to their first year (Van Engen et al., 2010).

The speakers chosen from Wildcat as model talkers were matched as closely as possible to the accents used in Baese-Berk et al. (2013). The selected talkers were all males with mid-range accentedness\(^2\), as rated by 50 native speakers (Van Engen et al. 2010). The accents selected were Korean, Mandarin, Thai, Farsi, Italian, and Russian. Table 1 shows the breakdown of accents used in Baese-Berk et al. (2013), compared with the model talkers analysed in Laturnus (Submitted) with their scores for accentedness.

\(^2\) Talkers in Wildcat are rated for accentedness, rather than intelligibility like the voices in NUFAESD used in Baese-Berk et al. (2013).
Table 1. L1s of model talkers used in Baese-Berk et al. (2013: NUFAESD), and in Laturnus (Submitted: Wildcat). Accentedness ratings are between 1 (no foreign accent) and 9 (strong foreign accent) in the corpus.

The Wildcat corpus consists of 60 scripted sentences: 30 ending in a highly predictable final word, and 30 ending in an unpredictable final word. Examples of each type are provided in (1).

1) For dessert he had apple pie. Predictable
   The color of a lemon is yellow. Predictable
   Mom talked about the pie. Unpredictable
   Mom thinks that it is yellow. Unpredictable

The variables chosen for analysis were Voice Onset Time (VOT) of word-initial voiced and voiceless stops, stressed vowels and their duration, and unstressed vowels and their duration. These were chosen because they were specifically suggested to be a possible source of systematic variation, and have been identified by numerous studies in the L2 literature as salient and measurably differing from English for speakers of varying language backgrounds (e.g. vowel contrasts: Barry, Hoequist, & Nolan, 1989; Flege et al., 1999; Fox et al., 1995; Oh et al., 2011; Tsukada, 2009; Wade et al., 2007; Wayland, 1997; unstressed vowels: Baker et al., 2011; Gut,
Based on the Systematic Variability Hypothesis, it would be expected that acoustic comparisons between speakers would reveal similarities in their productions of English. Laturnus (Submitted) found that data for each measure showed substantial variability across speakers, reflecting phonetic transfer from individual L1s, as well as substantial inconsistency and variability in pronunciation. No evidence was found for close acoustic overlap or similarity between speakers, contrary to a strong interpretation of this account.

The Category Loosening Hypothesis also predicts acoustic similarity between speakers, however it allows for generalization to occur via overlap between incoming speech and the aggregate of past experiences, rather than via a close acoustic match between the current speaker and a familiar voice. For example, new talkers’ /e/ tokens need not match closely with a single previous talker’s /e/ category, but instead could be similar to individual /e/ tokens of several different speakers. This allows for more variability in the input. The variability between individual talkers analyzed in Laturnus (Submitted) was large enough that some amount of overlap occurred between most talkers’ categories, though the extent of overlap varied, and was entirely absent in some cases. It remains an open question how much acoustic similarity is necessary for generalization under a Category Loosening account (i.e. does incoming variation need to fit completely within loosened boundaries based on past experience, or is partial overlap enough?).

Unlike these two explanations, the Weak Systematic Variability Hypothesis does not require acoustic similarity between talkers, instead tracking more general patterns in production.
errors and strategies. The data for the six talkers analyzed was consistent with this hypothesis for generalization, since similarity of this nature was observed between speakers. Four of the six talkers, for example, lacked a clear tense/lax contrast, none of them consistently reduced unstressed vowels, and all non-native speakers were significantly more variable than the American English model talkers. In order to explain generalization effects, the Category Loosening Hypothesis requires a novel accented talker to overlap in their productions with the aggregate of what a listener has heard before for each category. The Weak Systematic Variability Hypothesis instead requires the novel accented talker to differ from native speakers in his productions by making errors on the same structures as other non-native talkers the listener has heard before, without the requirement of acoustic similarity to previously heard speech. The next section will provide an overview of a framework that unites aspects of each of these hypotheses to provide an account of generalization that is consistent with the available data, both in the current body of work and in the literature.

1.3.4 Unifying hypotheses with the Ideal Adapter Framework

The Ideal Adapter Framework (Kleinschmidt & Jaeger, 2015) is a model of speech recognition that proposes listeners apply past experience to their current situation (e.g. their current interlocutor, location, or listening context) in order to limit the amount of statistical learning the perceptual system has to engage in at any given time. This is accomplished through three key objectives: recognizing when current situations are similar to past experiences, using those past experiences to generalize to new situations, and adapting to these new situations when necessary. Kleinschmidt and Jaeger (2015) outline a detailed computational model to account for a wide range of effects that have been documented in the literature, including perceptual adaptation at a
segmental level and greater speed and accuracy with familiar voices. Their account of adaptation
to novel talkers and generalization across groups of talkers is most relevant to the current work,
and so will briefly be outlined here.

The Ideal Adapter Framework (IAF) proposes that the speech perception system uses
statistical learning to track the distribution of incoming acoustic cue-to-category mappings and
build talker-specific models. More general models additionally exist in this framework (e.g.
*English speakers, males, or Canadians*), made up of the patterns that can be observed across
individual talkers of a group. These more general models are constructed at the point when the
perceptual system recognizes some commonality across multiple speakers, and the listener
recognizes some motivation for grouping them (i.e. efficiency of processing). In hearing a given
talker, the listener applies whichever model she believes is most relevant to the present situation.
If she recognizes the voice, either through top-down information (e.g. visually recognizing her
brother) or bottom-up (e.g. recognizing that the voice she’s hearing matches her brother’s voice),
she will apply a talker-specific model to the incoming speech stream.

The amount of perceptual adaptation to the voice that will occur will be based on the
listener’s confidence in 1) the accuracy of her talker-specific model, likely based on the amount of
input the model is based on, and 2) how applicable that model is to the current situation (e.g. if she
can see that her brother is talking, her certainty might be higher than if she only hears his voice).
More confidence leads to less perceptual learning, though some learning may always be necessary,
as individual cues are added to the model. Not having to extensively adapt to an interlocutor
constitutes a “head start”, because less statistical learning is required. If the most specific model
the listener feels confident applying to the current situation is more general (e.g. *Canadians*), more
perceptual learning will be necessary to build a talker-specific model for the present interlocutor. Kleinschmidt and Jaeger (2015) note that evidence for storing talker-specific information in memory in this way comes from studies reporting listeners are faster and more accurate in recognizing a familiar talker than an unfamiliar one (e.g. Eisner & McQueen, 2005; Palmeri et al., 1993). They also emphasize that listeners don’t have access to the real distribution of speakers’ categories, since their models are based on limited input. Speech recognition therefore involves various levels of uncertainty and the beliefs of the listener as they encounter new speech input.

Kleinschmidt and Jaeger suggest accent attunement and generalization occur in much the same way with non-native speech as with native speech. Generalization across groups of talkers is proposed to occur, as described, once properties are identified that are consistent between different people. Despite the immense variability of speech, the way categories vary across talkers isn’t completely random – there is structure, in that talkers can cluster along many variables, such as gender, age, dialect, identity, or non-native status in a certain language. By recognizing the patterns that co-occur between different members of these groups, the system is able to increase efficiency, since applying more specific models means less adaptation is necessary. Applying this framework to attunement studies, individuals might reach a point in their exposure where they are able to recognize commonalities, or systematic variation, in the speech of different talkers\(^3\). Depending on the nature of the input, these commonalities might be close acoustic matches for different categories (e.g. Xie & Myers' (2017) finding of acoustic similarity based on highly consistent,

\(^3\) Kleinschmidt and Jaeger (2015) do not provide an explicit account of how generalization to novel accented voices or novel accents more generally might occur beyond the suggestion of learning commonalities across talkers. The interpretation of how their framework might be applied that is presented here is my own.
single word stimuli), similarity based on the aggregate of past experiences with individual
categories (i.e. the Category Loosening Hypothesis), or more general patterns in how speakers
implement their categories (i.e. the Weak Systematic Variability Hypothesis: e.g. a lack of certain
contrasts, greater overall variability, or unreduced unstressed vowels). By building a model based
on this systematic variation, listeners would be able to apply the model to an unfamiliar talker they
believed to be similar enough to this category to benefit from beginning adaptation at that point,
based on those observed patterns.

Kleinschmidt and Jaeger note that the disadvantage of applying a highly specific model is
less flexibility: if the wrong model is selected, it takes a lot more speech input to adapt to the
current situation, based on a misleading or incorrect starting point. This observation can be applied
to the results of previous literature that suggest expectations for the signal can affect perception
and comprehension. The participants in Walker and Hay's (2011) experiment experienced a
perceptual advantage when their expectations for talker age and vocabulary were congruent with
the stimuli they were presented: older words heard in older voices were more quickly and
accurately responded to. In Niedzielski's (1999) vowel matching study in Detroit, when
participants were led to believe they were hearing a Canadian speaking, they selected vowels on a
continuum that were more raised, and therefore more consistent with stereotyped Canadian speech,
than those who believed they were hearing a fellow Detroiter. This was despite the model talker
being the same in both conditions. By selecting, for example, the Detroit model in this latter case,
participants began adaptation to the speech signal with a different set of cue-to-category mappings
than they would if they had selected the Canadian model, and are therefore biased towards hearing
vowels they believe to be more consistent with Detroit’s dialect. There are two possible ways
models could affect perception in this case. The first is that listeners’ own beliefs about what they’re hearing are used to interpret the speech signal, such that in hearing Detroiter's speech, these participants might perceive the input as consistent with what they think Detroiter sounds like, rather than noticing the raising that’s actually present. The alternative explanation is that listeners choose the wrong model, the General American model, where there is no Canadian raising, and apply it to the stimuli. Because they’ve chosen the wrong model, they experience slower processing, as they must readapt to what they’re hearing. Kleinschmidt and Jaeger don’t specify which explanation is more consistent with their framework, however elements of both accounts seem to be impacting perception. The participants in Niedzielski (1991) ultimately categorized the stimuli based on what they believed the Detroiter to sound like, suggesting their perception was biased based on the model they selected, however reaction time was not measured in this experiment, so it is possible they were also slower to react to the stimuli that did not match the model selected.

Babel and Russell's (2015) study provides a clearer example of both explanations impacting perception. They report a processing cost for listeners whose expectations for hearing an L2 accented voice did not match the acoustic input when they saw an Asian Canadian’s face and heard a native English speaker’s voice. In this situation, participants may have selected their L2 accented English model upon seeing the Asian Canadian faces, such that they then required more speech input in order to adapt to the native English model talker they were hearing, since the model did not match the input. This made them significantly slower and less accurate in transcribing the speech of the native English model talker. However, participants also rated the native talker as more accented when they saw these pictures than when the picture was of a white
person or when there was no picture, suggesting their perception of the signal was warped, or filtered, by their expectations. Expectations may therefore affect listeners’ processing in different ways: choosing the wrong model for the input might slow processing because more adaptation is necessary to reach adequate comprehension, and also skew perception, as listeners begin adaptation at a point where they already believe the input to sound a certain way (e.g. foreign-accented), based on past experience.

The IAF, when interpreted the way it is presented here, is able to account for adaptation and generalization across different L2-accented talkers by positing that listeners will reach a point of exposure to non-native speech where the system will identify enough patterns in pronunciation across different talkers to build a corresponding model. Modelling speech perception in this way accounts for key findings reported in the literature that bear on attunement and generalization. These include the role of top-down information, through sentential context and lexical disambiguation, and bottom-up information in perceptual adaptation (Kleinschmidt & Jaeger, 2015:14). Research suggests sentences play a greater facilitating effect in generalization between L2-accented speakers (Clarke & Garrett, 2004; Wade et al., 2007; Weil, 2001), and the suggested explanation is a benefit for greater context and lexical information constraining the potential input. The IAF additionally includes top-down information in the form of a listener’s certainty about their present situation as a component of the framework, such that visually recognizing an interlocutor (top-down information) will lead to higher certainty about the model being applied.

The Ideal Adapter Framework encodes talker-specific information while also allowing for abstraction, tracking commonalities across different linguistic categories or across different speakers. This allows listeners to both recognize familiar voices more quickly and accurately than
unfamiliar ones, and also generalize to previously unheard words, sentences, talkers, or accents based on prior experience. By incorporating certainty, perceptual learning is still a part of the system, which permits flexibility in adapting to new situations, however listeners need not re-adapt to familiar talkers at every meeting.

1.4 The effects of bias, exposure, and input variation in accent attunement

Using the Ideal Adapter Framework, the current dissertation addresses three questions that arise from the finding that listeners are able to generalize in an accent-independent way after training on several different speakers of unrelated L2 accents (Baese-Berk, Bradlow, & Wright, 2013). Methods in the three experiments designed to address these questions are as consistent as possible to allow for comparisons across groups.

1.4.1 Experiment 1: bias, adaptation and generalization across speakers

High variability training on multiple different voices has well-documented benefits in learning second-language contrasts (Bradlow et al., 1997; Lively et al., 1993; Wang et al., 2003, 1999), and in adaptation and generalization to novel accented speakers (Bradlow & Bent, 2008; Clopper & Pisoni, 2004; Sidaras et al., 2009). A more novel finding, however, is that training listeners on different talkers of unrelated accents facilitates comprehension of a new, untrained non-native accent. The first experiment in this dissertation sets out to replicate this result from Baese-Berk et al. (2013), while making changes to the methodology to examine the robustness of the findings.

In addition to testing for generalization after training, Experiment 1 investigates the potential effect of listener biases on perceptual adaptation to non-native speech. As discussed above, listeners’ expectations about the speech signal can influence their perception (e.g. Babel &
Russell, 2015). Though listeners are able to adapt to unfamiliar accents, a great deal of stigma exists surrounding accented English, and research suggests it may hinder individuals’ ability to communicate efficiently with non-native talkers (Yi et al., 2013, 2014). Sociolinguistic research on non-native speech has found that foreign-accented varieties are often associated with lower socioeconomic status and are often rated less favorably on traits like intelligence, competence and friendliness, by both native and non-native speakers (e.g. Bauman, 2013; Bouchard et al., 1977; Cargile, 1997; Cargile & Giles, 1998; Mulac et al., 1974; Ryan & Bulik, 1982; Ryan & Sebastian, 1980). By incorporating a measurement of listeners’ implicit biases towards non-native accented speech, this dissertation asks whether listeners with more negative biases perform worse on an attunement task than those with more positive biases. That is, is the perceptual system able to adequately adapt to the non-native input to facilitate generalization to novel talkers, despite biases? Or do these biases hinder learning?

1.4.2 Experiment 2: does real-world experience with L2 speech facilitate generalization?

Experiment 2 extends the investigation of Experiment 1 by moving adaptation beyond the laboratory. The findings of Baese-Berk et al. (2013) suggest listeners with more extensive experience listening to multiple different non-native accents might have an advantage processing non-native speech in general. Attunement studies, however, typically restrict participation to individuals who have limited prior exposure to non-native speech (e.g. Baese-Berk et al., 2013; Bradlow & Bent, 2008). The purpose of this is to ensure any effects observed are due to the training on non-native speech that is typically provided over the course of the experiment, however it prevents the investigation of more natural training people might attain through normal conversations. Experiment 2 therefore sets out to investigate whether these generalization effects
can also be observed when participants are not provided with training in the lab, but instead get
their training from the outside world, through interactions with non-native interlocutors over the
course of their lifetime. In this study, a group of speakers who self-report having extensive
experience with non-native talkers are contrasted with speakers who report that they have had little
exposure to non-native talkers over their lifetimes.

1.4.3 Experiment 3: what kind of variability is the right kind for generalization?

Baese-Berk et al. (2013), along with the experiments in Chapters 1 and 2, demonstrate that
training on non-native speech facilitates attunement and generalization. However, these studies are
not able to provide insight into what mechanism might explain why this effect occurs. Experiment
3, in conjunction with Laturnus (Submitted), is designed to further investigate this question. Base-
Berk et al. (2013) suggest their listeners had better comprehension of novel accented talkers
because they received training on different accents, and so were able to extract accent-independent
patterns from the input that facilitated generalization. Their discussion of the result signaled
support for the Systematic Variability Hypothesis. The goal of Experiment 3 is to test whether
exposure to L2 speech specifically is necessary for improved comprehension of novel L2 talkers,
or whether exposure to many native varieties of English, which can be very different in their
phonetic implementation, is sufficient to facilitate generalization.

The Category Loosening Hypothesis, as described, suggests listeners expand their category
boundaries in accordance with what they hear. Since accented talkers with different language
backgrounds will differ in their productions of English, this account suggests generalization occurs
because new input is likely to fall within the boundaries of previously-heard speech. Note that
while this would require exposure to variation in general to be a plausible account, since
boundaries would need to be widened enough for new variation to fall within them, the requirements for the Category Loosening hypothesis should be satisfied if there is only variability between talkers, rather than both between and within, as is characteristic for non-native speech. Experiment 3 therefore sets out to investigate the necessity of L2 training in generalization by training listeners on five different native regional English accents and testing them on two non-native speakers. If the Category Loosening Hypothesis is accurate, listeners should benefit from exposure to regional accents (e.g. New Zealand, Indian, Jamaican, etc), as it will provide them with greater variability between speakers, which will expand their categories. Because the model talkers in training are native speakers, they should have less within-talker variation than L2 speakers, but should still vary substantially from one another, being speakers of different native dialects. If no evidence of generalization is observed, however, it may be that listeners require exposure to L2 speech specifically. Experience with between-speaker variation may be insufficient because it doesn’t allow listeners to learn the patterns common to L2 speech, and such knowledge may be necessary for generalization to unfamiliar non-native voices, as is suggested by the Weak Systematic Variability Hypothesis.
Chapter 2: Implicit bias weakens accent attunement

In North America, accented non-native speakers face discrimination in the workplace, in education, in hiring, in housing applications, and in the criminal justice system (e.g. Lippi-Green, 1997). However, previous research on accent attunement, as discussed above, has found individuals are able to quickly and robustly adapt to non-native speech through exposure in the laboratory (e.g. Baese-Berk et al., 2013; Bradlow & Bent, 2008; Clarke & Garrett, 2004; Sidaras et al., 2009).

Despite the seeming ease of adaptation to L2 accents, however, native speakers still often hold negative opinions of non-native speakers. Matched guise studies, pioneered by Lambert, Hodgson, Gardner, and Fillenbaum (1960), are designed to reveal people’s implicit biases for or against different dialects. The task presents recordings to listeners spoken in multiple different accents or dialects (in a ‘true’ matched guise, these being recorded by the same multidialectal speaker) and asks them to judge the speaker on a number of traits including those mentioned above. Rubin and Smith (1990), for example, investigated undergraduates’ biases against non-native English-speaking teaching assistants (TAs) by having a Chinese-accented graduate student record lectures as moderately or heavily accented. While listening to a four-minute clip of the lecture, participants saw a picture of either a Caucasian or Asian “instructor” projected at the front of the room. The undergraduates gave lower teaching ratings for TAs they perceived as being more accented, regardless of whether the recording was more accented or not. On a survey about their experience with non-native TAs in the past, 42% of students indicated that they had dropped at
least one class because the TA was a non-native speaker, and 57% believed their grade had been hurt because the TA was non-native.

Another method for investigating individuals’ implicit biases is through an Implicit Association Test (IAT). First developed by Greenwald et al. (1998), the IAT measures an individual’s implicit bias by tracking their reaction time in aligning a concept (e.g. flower or insect) with an attribute (e.g. good or bad). Subjects are asked to sort stimuli (e.g. audio clips, pictures, or orthographically presented words) into categories on the left or right side of their monitor using keystrokes. During test trials, categories are combined on either side of the screen in either congruent (e.g. flower and good, insect and bad) or incongruent pairs (e.g. insect and good, flower and bad). The key concept is that people will find it easier to categorize stimuli when labels are congruent, and will therefore complete those trials more quickly (if someone has a pro-flower bias, they will be faster at categorizing images of flowers when flower and good appear on the same side of the screen than when flower and bad do).

In sociolinguistics and psychology, the IAT is a useful tool for measuring people’s biases toward socially meaningful categories or variables. Multiple studies, for example, report a stronger tendency among North Americans for associating non-Caucasian, and in particular, Asian faces with foreign and Caucasian faces with American (Devos & Banaji, 2005; Yi, 2013; Yi et al., 2013). Yi et al. (2013) found those with stronger Asian-foreign biases rated Korean-accented voices as more accented in an audiovisual condition than in an audio-only condition. Yi et al. (2014) investigated how listener bias to these same Asian faces affected speech perception using fMRI. They found that seeing the face of the Korean-accented talker exaggerated how accented participants perceived the voice to be, and also that greater Asian-foreign bias was associated with
decreased neural efficiency in the audiovisual condition for the accented voices, suggesting additional processing resources were needed. This body of research suggests individuals with stronger biases may also be less adept at attuning to non-native speech, though this remains an open question that the present work aims to empirically address using audio stimuli.

Though the stimuli in an IAT are typically visual, some studies have been done using audio clips. Pantos and Perkins (2013) used an IAT with auditory stimuli to compare implicit attitudes toward foreign-accented speech to explicit, self-reported attitudes. They found implicit biases favoured American English, while self-reported responses had a positive foreign-accent bias, ranking a foreign-accented talker as more reliable than a comparable speaker of General American English.

Campbell-Kibler (2012) used single-word audio stimuli in an IAT for an indexical investigation of the (ING) variable ([ɪn] vs. [ɪŋ] word-finally). She found participants implicitly associated [ɪn] with Southern states and [ɪŋ] with Northern states. Carter and Castellano (2016) used American city names produced in English and Spanish phonology as stimuli in an IAT study, finding participants were biased in associating English pronunciations with good, even when their first language was Spanish.

The current experiment builds on this body of literature by pairing an Implicit Association Test with a traditional accent attunement training paradigm. As described above, participants in such experiments receive laboratory training on non-native accented English via a transcription task, in which they listen to scripted sentences and type out what they hear. During test blocks, participants’ accuracy in transcribing unfamiliar voices and accents is taken as a measure of their attunement and generalization – how much the training on accented speech facilitated their
comprehension of the test voices. The current experiment is based closely on that of Baese-Berk, Bradlow and Wright (2013), who trained listeners on a Mandarin-accented speaker and four other speakers with different L1s. They found comprehension was facilitated at test for both a novel Mandarin-accented speaker and for a speaker of Slovakian, an accent not heard in training. This was in contrast to Bradlow and Bent (2008), whose participants performed no better on the Slovakian-accented talker at test than did the control participants, despite receiving training on multiple Mandarin-accented voices. The current work aims to replicate the pattern seen in Baese-Berk et al. (2013) while also investigating the effect of implicit biases towards non-native speech on participants’ ability to attune.

2 Experiment 1

2.1 Methods

The experiment consisted of an attunement task, wherein participants heard recorded sentences and orthographically transcribed them, followed by an Implicit Association Test (Greenwald et al., 1998), a social network questionnaire, and the Peabody Picture Vocabulary Test (PPVT). Each component is discussed further below.

2.1.1 Participants

Forty-three monolingual participants with no diagnosed speech or hearing disorders participated in this study for $15. The experiment took an average of 45 minutes to complete. 32 participants were recruited in New York City who self-reported having limited experience
listening to non-native speech. They were not born or raised in a major metropolitan area\(^4\) and had lived in New York City for less than one year. An additional 11 individuals were recruited from a college in upstate New York, also self-reporting limited exposure to non-native speech and meeting the same criteria. The results of these forty-three participants were compared to the low-exposure lifetime group in Chapter 3, who were from the aforementioned upstate New York college. The group of eleven individuals from upstate in this study therefore served to validate the results of the experiment, ruling out the possibility that differences between New York City participants and the control group were due to location or other factors aside from the experimental manipulation.

All participants completed a social network questionnaire, discussed below, to verify whether their reported personal contact with non-native speakers was consistent with low-exposure. The results of the questionnaire are discussed in §3.1. Two participants were excluded from data analyses upon review of their social network questionnaire and demographic survey. Both individuals had been born and raised in New York City, both had receptive vocabulary scores that were much lower than other participants (discussed further in §3.4), and both reported substantial exposure to Spanish-accented English.

\(^4\) Major metropolitan cities are defined here as populations over 100,000 people, as this corresponds the most clearly to having more than 25% of the city’s population being non-native speakers of English, according to 2011 census data (Gambino, Acosta, & Grieco, 2014).
2.1.2 Materials and procedure

2.1.2.1 Attunement task

As mentioned above, the design of this experiment is based on that of Baese-Berk et al. (2013) and Bradlow and Bent (2008). The attunement task consisted of 10 training blocks, followed by two test blocks. The model talkers were those analyzed in Laturnus (Submitted), discussed in Chapter 1. A second Mandarin-accented talker was selected from the Wildcat Corpus, also male with mid-range accentedness (Van Engen et al. 2010). The voices used during training were Korean, Mandarin, Thai, Farsi, and Italian. The test blocks consisted of the second Mandarin-accented voice and the Russian-accented voice. Table 2 shows the breakdown of accents used in Baese-Berk et al. (2013), compared with the model talkers in this experiment with their scores for accentedness.

<table>
<thead>
<tr>
<th>NUFAESD</th>
<th>Wildcat</th>
<th>Wildcat Accentedness Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korean</td>
<td>Korean</td>
<td>5.6</td>
</tr>
<tr>
<td>Mandarin</td>
<td>Mandarin</td>
<td>5.7</td>
</tr>
<tr>
<td>Thai</td>
<td>Thai</td>
<td>7.2</td>
</tr>
<tr>
<td>Hindi</td>
<td>Farsi</td>
<td>5.6</td>
</tr>
<tr>
<td>Romanian</td>
<td>Italian</td>
<td>7.0</td>
</tr>
<tr>
<td>Mandarin#2</td>
<td>Mandarin#2</td>
<td>5.8</td>
</tr>
<tr>
<td>Slovakian</td>
<td>Russian</td>
<td>5.7</td>
</tr>
</tbody>
</table>

Table 2. L1s of model talkers used in Baese-Berk et al. (2013: NUFAESD), and in this study (Wildcat). Accentedness ratings are between 1 (no foreign accent) and 9 (strong foreign accent) in the corpus.

As mentioned in Chapter 1, the Wildcat corpus consists of 60 scripted sentences: 30 ending in a highly predictable final word, and 30 ending in an unpredictable final word. Examples of each
type are provided again in (2) (from (1), above). These 60 sentences were broken into four lists of 15: each a randomized mix of high- and low-predictability sentences. All sentences were embedded in white noise at a signal-to-noise ratio of +5 dB, with noise extending 500 ms before and after each sentence, following Baese-Berk et al. (2013).

2) For dessert he had apple pie. Predictable
   The color of a lemon is yellow. Predictable
   Mom talked about the pie. Unpredictable
   Mom thinks that it is yellow. Unpredictable

Unlike in Baese-Berk et al. (2013), where training was split into two sessions occurring on consecutive days, all ten training blocks in this experiment took place in a single session, followed immediately by the two test blocks. Training was split into two sets, with a different list of sentences heard in each set.

Participants heard each model talker produce each sentence, meaning participants were trained on 5 repetitions of 15 sentences in each of 2 lists for a total of 150 training sentences. During both training and test blocks, talkers were randomized, as was the order of sentences.

A new list of 15 sentences was presented in each test block, with the sentence order and voice randomized by participant. Participants therefore heard 15 untrained sentences from each of the Russian and Mandarin speaker in a random order. Figure 1 illustrates the design of the current experiment compared to Baese-Berk et al.’s (2013) study.
Figure 1. A visualization of the Baese-Berk et al. (2013) study (left) compared to the current experiment (right). Mixed training blocks in the replication consist of 15 sentences produced by speakers of Korean, Mandarin, Farsi, Italian, and Thai, randomized within each block. Listeners hear 30 sentences from each talker over the course of the training session.

The attunement task and subsequent social network questionnaire were presented on a Dell Inspiron 1525 laptop running PsychoPy v.1.85.1 (Peirce, 2009) on Linux Mint. The laptop rested at eyelevel atop a Roost Laptop Stand. Responses were entered via a wired keyboard and mouse. AKG k240 studio headphones presented stimuli at a comfortable listening level.

No feedback was provided during training or test, which required orthographic transcription of each sentence. Participants heard each stimulus only once and began transcribing as soon as it started playing. Upon completion of one sentence’s transcription, participants pressed
the “Enter” key to hear the next sentence. They were unable to change their responses once the “Enter” key had been pressed.

To measure accuracy, keywords were chosen for each sentence. There were 102 keywords in total, which were all content words: 51 from high-predictability sentences, 51 from low-predictability sentences. Keywords varied in position within the sentence, with some being sentence-initial, some sentence-medial, and some sentence-final. Accuracy was calculated as the proportion of words recognized by scoring keywords in each sentence as correct or incorrect.

Performance on the test blocks was compared to the low-exposure condition described in Chapter 3. Instead of being trained on five non-native accents, this control group heard five male, native speakers of American English during training. These voices were selected from the Wildcat Corpus for their homogeneity of dialect and reading style, as judged by the author and two other trained phoneticians. Participants completed the same test blocks as in the present study, randomized by sentence order and model talker. Aside from the training voices, this control group and the current experiment were identical in methods and set up. Comparing the two groups therefore permits the investigation of the effect of non-native training on generalization to 1) a novel talker of an accent heard in training (i.e. Mandarin), and 2) a novel talker of an untrained accent (i.e. Russian).

2.1.2.2 Implicit Association Test

Upon completion of the two test blocks, participants completed an Implicit Association Test (IAT), the results of which were compared to individuals’ performance on the attunement portion of the experiment. The IAT administered was adapted from the OpenIAT (Stafford & Scaife, 2015) for PsychoPy, available through the Open Science Framework (https://osf.io).
The IAT consists of five blocks of speeded categorization. The first two blocks provide training on how to appropriately respond to the stimuli. Visual stimuli appear in the form of either pictures or text, while auditory stimuli are presented over headphones.

The first block of the IAT trains participants on *attribute categories*, requiring them to sort words presented orthographically into labels that are associated with a different key or button on the input device (e.g. ‘negative’, ‘positive’). Positive attribute words were *good, joy, love,* and *pleasant,* while negative attribute words were *bad, agony, harm,* and *nasty.* Responses in the current study were logged via buttons 2 and 8 on a DirectIN High Speed Button-Box v2012. During the second block, participants learn the *concept categories* by sorting auditory stimuli into one of two categories (e.g. “Press 2 for ‘accented’, 8 for ‘unaccented’”). Unlike with attribute categories, concept categories in the current experiment require participants to sort based on features of the audio stimuli other than semantic meaning. For this reason, only semantically neutral phrases are used as audio stimuli (explained in detail below). Feedback is provided during these first two blocks to ensure participants learn how stimuli are expected to be categorized. Block three, the first test block, combines concept and attribute categories, such that one of each category is assigned to the same response key (e.g. “Press 2 for ‘accented’ or ‘positive’”). The purpose of co-assigning categories to the same key is to test how readily participants can associate L2-accented speech with positive attributes. Block four is the final training block. It is identical to Block 2, but with key assignments reversed (e.g. “Press 2 for ‘unaccented’”). Block five is the

---

The terms ‘accented’ and ‘unaccented’ were deemed to be less opaque to naïve participants than ‘native’ and ‘non-native’, and more accurate than ‘American’ and ‘foreign-accented’. That ‘accented’ was intended to refer to non-native accented speech was explained explicitly to participants prior to the start of the Implicit Association Test.
second test block and is identical to Block 3 but with key assignments for the concept categories reversed (e.g. “Press 2 for ‘unaccented’ or ‘positive’”). This is to test how readily participants associate L2-accented speech with negative attributes. Participants’ reaction times during the test blocks are used to calculate a score as a measure of implicit bias for non-native-accented speech relative to American English speech. The order of concept category labels on the left or right side of the screen is counterbalanced between participants. Figure 2 provides an example of what participants might see for a trial in Block 3 and Block 5.

![Figure 2. Examples of what was presented on screen during Blocks 3 and 5. Participants were required to quickly sort both audio clips and orthographic words, one at a time, in Blocks 3 and 5.](image)

For the auditory stimuli, two-word collocations were extracted from sentences used during training in the attunement task. These collocations were pre-tested for semantic neutrality using Amazon Mechanical Turk. Pantos and Perkins (2013) used multi-syllabic words (e.g. *probability*) and collocations (e.g. *assistance first*) extracted from another task in the same study, where participants listened to a fictional testimony from a Korean-accented actor and Philadelphia-accented actor portraying physicians in a medical malpractice trial. Stimuli were pre-tested for
semantic neutrality (rated in an online survey as bad-good, pleasant-unpleasant, and negative-positive). They were also equalized for amplitude. In line with Pantos and Perkins (2013), the current work elicited explicit, self-reported ratings for amplitude-equalized collocations from training sentences.

Fifty participants completed the online survey through Amazon Mechanical Turk (MTurk). Participants were paid $0.50 for completing the task, which took an average of three minutes. MTurk allows experimenters to restrict the task to potential participants with IP addresses in a specific geographic location. In the current study, location was restricted to the United States. Participants additionally had to confirm that they were native speakers of American English, that they had lived in the United States from birth until age 13, and that all of their parents or guardians spoke English and only English to them during those years.

Participants rated each of twenty-four multi-syllabic sequences extracted from the training sentences in the attunement task. Listeners heard each collocation read by a native speaker of American English and rated it along three different dimensions (negative-positive, bad-good, unpleasant-pleasant) on a scale from 1 (negative/bad/unpleasant) to 7 (positive/good/pleasant). A total of eight collocations were chosen that had an average rating of semantic neutrality (4 on the scale, with a standard deviation of 0.15). The ratings for the eight chosen stimuli on each attribute are provided in Table 3.
Table 3. Results of pre-screening for semantic neutrality for the eight chosen IAT stimuli.

Excerpts were rated on a scale from 1 (bad/negative/unpleasant) to 7 (good/positive/pleasant). 4 indicates a neutral rating. Average rating is an average across all three semantic judgements.

<table>
<thead>
<tr>
<th>Audio excerpt</th>
<th>Negative/Positive</th>
<th>Bad/Good</th>
<th>Unpleasant/Pleasant</th>
<th>Average rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>A watch</td>
<td>4.00</td>
<td>4.02</td>
<td>4.10</td>
<td>4.04</td>
</tr>
<tr>
<td>Dad talked</td>
<td>4.04</td>
<td>3.92</td>
<td>4.12</td>
<td>4.03</td>
</tr>
<tr>
<td>Her feet</td>
<td>3.81</td>
<td>3.85</td>
<td>3.81</td>
<td>3.82</td>
</tr>
<tr>
<td>Her head</td>
<td>4.12</td>
<td>4.13</td>
<td>4.00</td>
<td>4.08</td>
</tr>
<tr>
<td>The colour</td>
<td>3.90</td>
<td>3.96</td>
<td>4.08</td>
<td>3.98</td>
</tr>
<tr>
<td>Their nests</td>
<td>4.17</td>
<td>4.25</td>
<td>4.29</td>
<td>4.24</td>
</tr>
<tr>
<td>Twenty-eight</td>
<td>4.23</td>
<td>4.35</td>
<td>4.29</td>
<td>4.29</td>
</tr>
<tr>
<td>We talked</td>
<td>4.02</td>
<td>4.02</td>
<td>4.19</td>
<td>4.08</td>
</tr>
</tbody>
</table>

The goal of the IAT in this experiment is to get at participants’ biases against non-native speech specifically, rather than against particular nationalities or foreign people in general. Research suggests American listeners have biases against Mandarin (Bauman, 2013; Lindemann, 2005), Russian (Lindemann 2005), and Italian accents (Giles, 1970; Mulac, Hanley, & Prigge, 1974). Korean, Farsi, and Thai, on the other hand, have been found to be less familiar accents of English, with listeners unable to correctly identify the ethnicity or language background of model talkers (e.g. Lindemann, 2003) for Korean-accented talkers, (Eisenchlas & Tsurutani, 2011) for Farsi), or self-reporting a variety as unfamiliar (e.g. Lindemann 2005 found Thai speakers among the least familiar to participants out of a list of 58 countries with populations over two million). The voices used in the IAT were therefore these less-familiar accents of Korean, Farsi and Thai. During the training block of the IAT, described above, participants heard a male Turkish speaker and a fourth American English speaker reading the same pre-screened stimuli as the test voices.
Model talkers were selected from the Wildcat Corpus and were different than those heard during the attunement task. In addition to the non-native voices, participants heard the same tokens produced by three male, native General American English speakers. Each individual sentence fragment was heard 32 times by each listener: twice from each training talker in each training block (2*2*2 = 8), and twice from each test talker in each test block (2*6*2 = 24). Using Praat (Boersma & Weenink, 2017), all audio clips were equalized for amplitude and analyzed for pitch variation across talkers. Stimuli in the IAT were presented in the clear, rather than in noise.

Reaction times in Blocks 3 and 5 are used to calculate a $D$-score for each participant. This is done by subtracting mean reaction time for incongruent trials (i.e. ‘accented’ and positive) from mean reaction time for congruent trials (i.e. ‘accented’ and negative), divided by the standard deviation of incongruent and congruent trials. The equation is provided in (3). The resulting score will therefore be more negative when biases are more stereotypical (i.e. stronger bias for associating non-native accented speech with the label negative). This way of calculating the $D$-score is the opposite of what is typically done in the literature, where congruent trials are subtracted from incongruent trials and divided by the standard deviation. The calculation is reversed here so that more negative biases correspond to more stereotypical biases.

3) \[ \frac{\text{Mean congruent} - \text{mean incongruent}}{\text{Standard deviation}} \]

2.1.2.3 Social network questionnaire

The social network questionnaire, adapted from Stoessel (2002), was administered following the attunement task. The questions are reproduced here in (4-9) and are also provided in
the Appendix. Psychopy (Peirce, 2009) displayed each question in (4a) followed by the question in (4b). Participants answered all questions in (4) before proceeding to question (5).

4) a. How many [friends
coworkers
house or roommates
neighbors
family members
teaching assistants
professors] do you have/have you had who are non-native speakers of English?

b. What is their native language? Guess or be general if you are not sure. Type "n/a" if you don't know any non-native speakers in this category.

5) When you were in high school, how many friends did you have who were non-native speakers of English?

6) When you were young (elementary school), how many friends did you have who were non-native speakers of English?

7) In your high school, were at least 30% of students non-native speakers of English?\(^6\)

8) In your elementary school, were at least 30% of students non-native speakers of English?

9) Do you think 30% or more of the people you regularly interacted with (either close to you or in customer service positions) when you were younger (elementary or high school age) were non-native speakers of English?

\(^6\) 30% was chosen as the criteria in these questions to correspond to the criteria used in recruitment, where census data indicates metropolitan areas over 100,000 people have a substantial portion of the population being non-native English speakers (see footnote 4).
Results from analyses of accuracy by social network questionnaire responses in Chapter 3 suggest participants can be adequately considered low exposure if they have one family member or fewer with whom they interact regularly and if they answered “yes” to no more than one of the three questions in (7-9).

2.1.2.4 PPVT

The Peabody Picture Vocabulary Test-IV is a standardized assessment used to measure American English-speaking individuals' receptive vocabulary. The experimenter administered the test following completion of the social network questionnaire to test for differences between the experiment groups. Larger receptive vocabulary size has been shown to correlate with greater accuracy for comprehension of speech in noise among adult participants (e.g. Banks, Gowen, Munro, & Adank, 2015; Bent, Baese-Berk, Borrie, & McKee, 2016; Tamati, Gilbert, & Pisoni, 2013).

PPVT word lists are normed for age, so a participant’s age determines the item number at which their assessment begins. Four pictures per word are displayed on a computer screen. The experimenter reads out a given word and the participant selects the picture that best corresponds to the word’s meaning. Eight or more mistakes in one list finishes the assessment, with the participant completing the rest of the current list but not proceeding further. An individual's score is calculated by subtracting the number of mistakes from the total number of items administered.

2.1.2.5 Statistical modelling

Statistic analyses were conducted using the lme4 package (Bates et al. 2015) in R (R Core Team 2016). The addition of all aforementioned variables (i.e. condition, recruitment location, IAT $D$-score, PPVT score, test language, keyword position, and keyword predictability) in a
logistic mixed effects model of accuracy on keywords resulted in failed convergence. Consequently, a base model was constructed of accuracy with a random effect of participant. This model was then compared to subsequent models in a step-up manner using R’s `anova` function, in order to determine the significance of each fixed effect in improving model fit. None of position, predictability, or recruitment location significantly impacted model fit, regardless of reference level, so they were excluded from the maximal model. The addition of PPVT score prevented model convergence regardless of the stage it was added and so was assessed separately in a linear model of PPVT score by condition and recruitment location, to test for differences between participant groups in receptive vocabulary. The most complex model of accuracy that was able to converge was a logistic mixed effects model of keyword accuracy with fixed effects of training condition (i.e. L2-accented training or American English training), IAT D-score, and test language (i.e. Mandarin and Russian). Random intercepts were included for listener and sentence, and random slopes for test language by listener. Statistical output tables for each comparison are provided in the Appendix.

To further assess the effect of IAT D-score on accuracy within each participant group, additional logistic mixed effects models of accuracy by D-score were run within each training condition (i.e. training on non-native accents or American English accents) and within each college, with a random effect of participant.

2.2 Results

2.2.1 Social network questionnaire

Table 4 provides the average responses to each question in the social network questionnaire, by training condition. Results of analyses run in Chapter 3 found participants’
number of non-native friends and family, as well as how much of their community consisted of 30% or more non-native speakers (#6-8 on the questionnaire, above) significantly predicted accuracy on test trials. The model in Experiment 2 was run over all participants, half of whom had substantial prior exposure to non-native speech, while the other half were low-exposure. For participants in the current experiment, trained on L2 speech, a mixed-effects logistic regression model found none of the questions on the SNQ were significant predictors of accuracy. This is interpreted as being due to the homogeneity of participants’ prior exposure to non-native accented speech, as all participants were explicitly recruited as having low exposure. The only two participants here who diverged substantially from these averages were the two mentioned in §2.1, who each identified more than 10 family members and more than 10 friends who were non-native speakers of Spanish. Because of their greater prior exposure to non-native speech, they were excluded from further analyses.
### Table 4. Average number of individuals in each category of the social network questionnaire by condition.

<table>
<thead>
<tr>
<th>SNQ question (number of)</th>
<th>Experimental group average</th>
<th>Control group average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-native accents exposed to</td>
<td>2.86</td>
<td>2.91</td>
</tr>
<tr>
<td>Family members</td>
<td>0.3</td>
<td>0.29</td>
</tr>
<tr>
<td>Friends</td>
<td>1.54</td>
<td>2.1</td>
</tr>
<tr>
<td>Other (profs, TAs, neighbours, roommates)</td>
<td>4.46</td>
<td>4.7</td>
</tr>
<tr>
<td>Childhood friends (high school, elementary school)</td>
<td>1.6</td>
<td>2.7</td>
</tr>
<tr>
<td>Non-native accents with daily exposure</td>
<td>1.29</td>
<td>2.12</td>
</tr>
<tr>
<td>30% community (#6-8)</td>
<td>0.21</td>
<td>0.23</td>
</tr>
</tbody>
</table>

2.2.2 **Attunement task**

Neither position ($\chi^2(2)=0.17, p=0.92$), nor predictability ($\chi^2(1)=0.12, p=0.73$) significantly improved model fit. Test accent (i.e. Mandarin or Russian) did improve model fit ($\chi^2(1)=25.5, p<0.001$), however no significant effect of test accent was found in the maximal logistic mixed effects model that was run ($\beta=-0.7, z=-1.12, p=0.268$). A significant effect was found for IAT $D$-score as a predictor of accuracy ($\beta=0.96, z=3.68, p<0.001$), and for training condition ($\beta=0.57, z=3.01, p=0.003$). Participants who were trained on non-native speech ($\mu=96.10\%, \sigma=0.194$) were significantly more accurate than the control group ($\mu=93.93\%, \sigma=0.239$) on test trials. Figure 3 plots proportion accuracy on test trials by condition. IAT $D$-score as a predictor of accuracy is discussed further below, in section 2.2.3.
Figure 3. Proportion accuracy on test trials by training group (non-native-accented English or the American-accented English control)

Plots of errors by block reveal that participants trained on non-native accents did improve over time, which is consistent with previous research (Baese-Berk et al., 2013; Bradlow & Bent, 2008). Figure 4 shows the number of errors increases slightly in Block 6, which coincides with the beginning of the second set of sentences participants are trained on.
2.2.3 Implicit Association Test

Participants with more positive D-scores were more accurate in their transcriptions ($\beta=0.96$, $z=3.68$, $p<0.001$), regardless of whether they were in the accent-training ($\beta=0.88$, $z=2.53$, $p=0.011$) or control condition ($\beta=0.94$, $z=3.36$, $p<0.001$).

Logistic mixed effects models of accuracy by IAT score within each recruitment location found the same pattern, where participants with more positive (i.e. less stereotypical) implicit biases were more accurate on test trials (NYC participants: $\beta=0.99$, $z=2.13$, $p=0.033$; upstate participants: $\beta=0.82$, $z=3.46$, $p<0.001$).

Figure 4. Number of errors by training block for participants trained on non-native accents.
Figure 5 plots accuracy on test trials by IAT $D$-score\textsuperscript{7}. The curved line shows that accuracy increases as IAT score gets increasingly positive. Positive $D$-scores indicate a stronger implicit bias for categorizing non-native speech with the label ‘positive’. Figure 6 provides a clearer picture of the range of IAT scores plotted by accuracy.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{accuracy_iat.png}
\caption{Plot of accuracy on test trials by IAT score. Black dots at 0.00 and 1.00 represent each participant’s accuracy for each keyword. Positive $D$-scores indicate a stronger implicit bias for categorizing non-native speech with the label ‘positive’.}
\end{figure}

\textsuperscript{7} Vertical jitter has been added to the plot to more clearly indicate clusters in accuracy by IAT score.
Figure 6. Scatter plot of IAT scores by accuracy for trained participants (one IAT score per person)

Because IAT scores are calculated from participants’ reaction time on audio test trials, their reaction time on orthographic trials during the test blocks served as a control. To ensure the task was being completed as expected, accuracy on orthographic trials was compared to accuracy on audio trials during the test blocks. Figure 7 shows accuracy on orthographic and audio trials in the IAT for each participant. Note that the data for the orthographic analysis is collapsed over three experiments; the audio stimuli will be analyzed separately for each experiment in subsequent chapters. Average accuracy over all 143 participants was 96.63% for orthographic trials, indicating listeners were performing as expected on the task, finding these trials more easily categorizable than audio stimuli (μ = 89.92%). This is expected because the attribute category labels (‘positive’ and ‘negative’) stayed on the same sides of the screen throughout the task, while the labels used to
categorize the audio stimuli (‘accented’ and ‘unaccented’) switched sides of the screen during each test block.

Figure 7. Proportion accuracy on orthographic and audio IAT trials per participant.

2.2.4 PPVT

PPVT scores for each participant are included in the Appendix. Condition was not found to be a significant predictor of PPVT score in a linear model ($\beta=2.7$, $t=1.2$, $p=0.236$), nor was participants’ school ($\beta=-3.034$, $t=-1.35$, $p=0.181$). Receptive vocabulary trended towards
significance as a predictor of proportion accuracy on test trials over all participants ($\beta=0.0009$, $t=1.992$, $p=0.0502$).

2.3 Discussion

Experiment 1 aimed to replicate the results of Baese-Berk et al. (2013) while additionally modifying the methods, with the goal of making the task more difficult. Training and test occurred on the same day in this experiment, rather than having training spread over two consecutive days. Model talker and sentence order was additionally randomized in each block. Despite these changes, listeners who received training on five unrelated, non-native accents did have greater accuracy on test trials than those in the control group, who were trained on five native speakers of American English. This included improved comprehension of both an accent they had been trained on (i.e. Mandarin) and an accent they had not (i.e. Russian). The results replicate the effects reported in Baese-Berk et al. (2013), who found participants trained on five unrelated foreign accents had greater accuracy transcribing a new talker of an accent included in training and a new talker of an untrained accent.

Unlike Baese-Berk et al. (2013), however, the accuracy in this experiment for all participant groups was very high. Baese-Berk et al. report the average accuracy of their control participants (from Bradlow and Bent, 2008) being approximately 75%. In contrast, the control group in this experiment was 94%. The model talkers used here were from the Wildcat Corpus, and so were rated by native listeners for accentedness, while those in Baese-Berk et al.’s study were rated for intelligibility. Munro and Derwing (1995) found listeners’ ratings of accentedness relative to perceived comprehension and actual intelligibility were not highly correlated, with listeners rating accentedness harshly (i.e. more accented) despite high comprehension of the
speech they heard. It is therefore possible that the talkers chosen for this experiment, who all had mid-range accentedness scores, were in reality more intelligible for listeners than those used in Baese-Berk et al. (2013). This may have caused the higher accuracy scores in the current experiment (and all other experiments in this body of work, as Chapters 2 and 3 will demonstrate), despite having the same SNR. The topic of increased accuracy is returned to in Chapter 5.

No significant differences were found in receptive vocabulary score between either the control and accent-training condition, or between the two locations used for recruitment. It is therefore unlikely that participants in the accent-training condition had greater accuracy from an advantage due to vocabulary abilities. However, research has linked vocabulary size to greater comprehension of speech in noise (e.g. Banks et al., 2015; Bent et al., 2016; Tamati et al., 2013), and the analyses run over all participants and within the accent-training condition did find a pattern consistent with these results. This suggests some participants benefitted from greater receptive vocabulary size, however there is no evidence for their being unequally distributed between the control and accent-training condition.

The results reported here provide further evidence that listeners are able to improve their understanding of non-native speech with exposure. It is additionally consistent with literature on attunement that suggests hearing full sentences, rather than words in isolation, is beneficial for generalization. In Weil (2001), generalization occurred from one Marathi speaker to another Marathi speaker only during sentential tasks (and not at all to a Russian speaker). In comparison, Wade (2007) found no improvement in word recognition accuracy for listeners exposed to monosyllabic words produced by Spanish-accented speakers, even after three days of training, suggesting single-word utterances may be inadequate for attunement.
All accent-trained participants improved comprehension during training, including those with more negative implicit biases. This suggests education and training on accent attunement could be beneficial to a wider audience. Derwing, Rossiter, and Munro (2002) trained social work students on 1) cross-cultural awareness and 2) provided instruction on attitudes towards and comprehension of L2 speech. Participants in one condition had training of both types, another only received training on cross-cultural awareness, while a third group had no training. Participants completed a pre-test prior to training that required listening to a Vietnamese-accented English passage and transcribing sentences. Following training, all participants completed the same task again. The authors found all groups significantly improved comprehension of Vietnamese-accented speech, with no significant differences in accuracy between groups. Both experimental groups showed greater empathy for immigrants after training. However, those trained explicitly on Vietnamese-accented English had more confidence they could successfully interact with L2 speakers. The increase in accuracy of even the control group, who received no training or exposure beyond the pre- and post-test passages, is consistent with the hypothesis that only a small amount of exposure is needed for attunement. Listening to the pretest alone may have been sufficient to improve comprehension.

A more complicated result of the current study is the finding that individuals with stronger biases for categorizing non-native accented speech as negative were less accurate in their comprehension than those with stronger biases in the opposite direction, even after training. While studies such as this one suggest attunement can occur merely through exposure to L2-accented speech, other research has found overcoming biases to be a more difficult process, and that having internal motivation is more effective than external motivators (Gonsalkorale et al., 2011; Maddux,
Barden, Brewer, & Petty, 2005; Rudman, Ashmore, & Gary, 2001; Allen et al., 2010). Plant and Devine (1998) developed the Internal and External Motivation to Respond Without Prejudice Scales (IMS and EMS, respectively) to measure these sources of motivation in individuals, specifically in responding without racial prejudice. Statements on the IMS include, for example, “Being nonprejudiced toward Black people is important to my self-concept”, while statements on the EMS include “I attempt to appear nonprejudiced toward Black people in order to avoid disapproval from others.” Using these scales, Devine, Plant, Amodio, Harmon-Jones, and Vance (2002) conducted a series of experiments designed to investigate the effect of motivations on implicit and explicit bias measures. They found explicit bias was moderated by participants’ internal motivation to respond in an unprejudiced manner. People who had greater internal motivation and low external motivation to respond without prejudice showed lower implicit race bias than all other participants. This suggests, they argue, that the source of motivation plays a greater role than the amount of motivation in the effectiveness of responding without bias, since individuals who had both high internal and high external motivation did not show the lowest level of implicit bias.

To investigate the malleability of implicit biases, Rudman et al. (2001) measured implicit and explicit biases of students voluntarily enrolled in a prejudice and conflict seminar. Previous research suggests compulsory diversity training can have negative effects as a result of backlash (e.g. Brehm, 1966). With voluntary participants, however, Rudman et al. found both implicit and explicit biases had decreased by the end of the course, despite not differing significantly from the control on either measure prior to training. Given the nature of the experiment, wherein students participated through self-selection, it is impossible to tease apart the role of the course content in
changing participants’ minds versus simply the openness of the participants in having their minds changed.

The literature on motivations and biases centres largely around racial biases, rather than speech-related biases. However, as outlined in the introduction of this chapter, several studies have linked implicit biases about Asian faces to greater perceived accentedness. Race therefore appears to play a role in speech perception, suggesting findings in the racial bias literature may be relevant here. Individuals who have an internal desire to recognize their implicit biases and work to mitigate or eliminate them may therefore have some success through such things as prejudice and conflict resolution training.

While participants in this study with higher implicit bias scores toward non-native accented speech were less accurate in their transcription, all participants achieved accuracy above 80%, with most clustering above 90% (high accuracy results are discussed further in Chapter 5). Even those with more stereotypical biases were therefore able to understand the test talkers quite well. This calls into question the results reported in studies of students’ evaluations of non-native instructors in the classroom, where non-native speakers are often rated less favourably than their native counterparts, and where these ratings have been found to interact with ethnicity (Rubin, 1992; Rubin & Smith, 1990). Rubin (1992) conducted several experiments designed to elucidate nonlanguage factors influencing students’ judgements of their non-native teaching assistants. In one study, participants heard brief lectures while seeing a photo of the ‘instructor’, either a Chinese woman or a White woman, and were asked to complete a listening comprehension test. The lecture heard by all participants was recorded by a native speaker of American English. Following the listening test, participants rated the instructor on a variety of scales related to attitudes, background,
values, and appearance. Rubin found participants performed better on the listening test when the photo seen was the White instructor than when it was the Chinese instructor. The Chinese instructor was also rated as more accented and less standard than the White instructor, even though all participants heard Standard American English. In a second study, Rubin (1992) found students rated their TAs more negatively when they perceived the TA’s accent to be non-native, or ‘foreign’. He additionally found that students who had taken more classes with non-native instructors performed better in a listening comprehension test. Studies like these therefore suggest students’ ability to understand the speech of classroom instructors is not the key source of their dissatisfaction. Instead, accentedness seems to be used as a justification for stigma. Listeners place the onus of effective communication on L2 speakers, rather than acknowledging that their own negative attitudes and perception of L2-accented talkers are likely influenced by other factors, such as racial prejudice. Participants in Baese-Berk et al. (2013) and Bradlow and Bent’s (2008) control group achieved 75% accuracy, which is much better than chance, suggesting that even in more difficult listening environments, people are able to comprehend non-native speech fairly well. The amount of stigma non-native speakers face is therefore unwarranted if only comprehension is to blame. Future work should investigate implicit bias and comprehension of accented L2-speech in more naturalistic settings, like the classroom, while also testing for racial biases, in order to gain a clearer understanding of the factors influencing poor teaching ratings in higher education for non-native instructors.

2.4 Conclusion

This study investigated the effect of participants’ implicit biases toward L2 speech on their ability to comprehend unfamiliar accents and voices. Individuals who received training on five
unrelated non-native accents were significantly more accurate in their comprehension of an unfamiliar accent than were those trained on speakers of American English, which is consistent with prior work. Despite the widespread belief that L2-accented speech is difficult to understand and constitutes a barrier to communication, these data contribute to the body of literature demonstrating the ease with which processing costs associated with L2-accented speech may be overcome. As other scholars have suggested (e.g. Rubin & Smith 1990), education and training on accent attunement and the stigma faced by L2 speakers would be beneficial to students, teachers, and other professionals who encounter L2 speech on a regular basis in the workplace and interpersonally. The effect of implicit bias is less straightforward, however, as those with more negative biases towards non-native speech were less accurate in their comprehension of new voices and accents, even after training on L2-accented speech. However, overall accuracy even among these participants was very high, providing further evidence that native speakers’ complaints of comprehension difficulties towards L2 learners, as reported in the sociolinguistic and higher education literature, do not fit with actual abilities to understand non-native speech and are instead likely due to other factors.
Chapter 3: Greater lifetime exposure facilitates generalization

Given the findings presented in Chapter 2 that attunement occurs after listening to non-native talkers in a laboratory setting (see also Bradlow & Bent 2008, Sidaras et al. 2009, Clarke & Garrett 2004), one might expect significant exposure over the course of one’s life to have a similar effect in facilitating one’s ability to comprehend a novel accent. The results of Experiment 1 and Baese-Berk et al. (2013), in particular, suggest greater experience with a variety of unrelated non-native accents might improve comprehension of an unfamiliar accent encountered in the wild. Beyond the laboratory, Atagi and Bent (2016) found native English speakers from an area where South Koreans were the largest population of international students were able to accurately identify the model talkers’ language background, but were less accurate in identifying Spanish-accented speakers, a less common accent variety in the area. Witteman, Weber, and McQueen (2013) conducted a cross-modal priming experiment with Dutch listeners who either had limited prior exposure to German-accented Dutch, or extensive prior exposure. Participants heard words that varied in their level of accentedness. The authors found inexperienced listeners were primed only by weakly and medium-accented speech, while more experienced listeners were primed by all levels of accentedness. In an analogous situation to accent comprehension, McGarr (1983) found that teachers and other individuals who worked frequently with deaf students were consistently better at recognizing words and sentences produced by deaf children than were inexperienced listeners, likely because of their experience processing this kind of variation in the speech signal.

Experiments on attunement, however, often screen participants for significant prior exposure to non-native speech (e.g. Baese-Berk et al. 2013, Bradlow & Bent 2008). While this
allows researchers to mitigate differences in participants’ language backgrounds and isolate the
effect of training during the course of the experiment, it does not permit the investigation of
whether real-world exposure to non-native speech shows similar generalizability to novel accents.
Porretta, Kyröläinen, and Tucker (2015) found participants with greater experience listening to
Chinese-accented English showed stronger identity-priming effects for Chinese-accented words,
as well as faster word recognition in a visual world eye-tracking experiment. Their results provide
further evidence that experience with a particular accent can cause listeners to develop speaker-
independent adaptation to that particular accent. By comparing populations that differ in their level
of real-world exposure to multiple non-native (L2) accents, this experiment examines whether
accent-general effects found in attunement studies are mirrored in the wider population, who must
attune to new speakers in everyday life through exposure outside the laboratory. The results
presented below confirm that experienced listeners are more accurate in transcribing non-native
accented speech.

3 Experiment 2

3.1 Methods

The experiment consisted of an attunement task, wherein participants heard recorded
sentences and orthographically transcribed them, followed by an Implicit Association Test, a social
network questionnaire, the Peabody Picture Vocabulary Test (PPVT), and a demographic survey.
With the exception of the attunement task, each of these components was identical to those
described in Experiment 1 (Chapter 2). The attunement task is explained in detail below.
3.1.1 Participants

Seventy monolingual participants with no speech or hearing disorders participated in this study for $15. The experiment took between 45 minutes and one hour to complete.

24 participants were recruited in New York City who had been born and raised in major metropolitan cities and had been living in New York City for at least two years at the time of participation. They self-identified as having substantial experience listening to non-native speech. An additional 38 individuals from a college in upstate New York participated in the experiment, all self-reporting limited exposure to non-native speech at the time of recruitment. This location was chosen to mitigate the chance of participants having substantial daily exposure to non-native speech. The demographics of the college are reported as being 64% White, 15% Hispanic/Latino, 5% African American, and 16% other, with 95% of enrollment being in-state, 4% out-of-state, and 1% of the total 7,692 students coming from foreign countries (National Center for Education Statistic, 2014). Students recruited from this college were not born, raised, or currently living in major metropolitan areas. All participants completed a social network questionnaire, discussed below, to verify whether their reported personal contact with non-native speakers was consistent with their categorization as high- or low-exposure. Seven of the upstate participants were re-categorized as high-exposure after evaluation of their questionnaire, such that analyses were ultimately conducted on 31 participants in the high-exposure group and 31 participants in the low-exposure group. The results of the questionnaire are discussed in §2.2.1.

An additional eight monolingual students with no language or hearing disorders were recruited from the aforementioned college in upstate New York, all identifying as having low lifetime exposure to non-native speech. These participants completed the attunement task without
training on speech in noise, in order to evaluate the effect of training with speech in noise on accuracy at test with non-native speech recognition.

3.1.2 Materials and procedure

3.1.2.1 Attunement task

The design of this experiment was identical in structure to that of Experiment 1. The attunement portion for the trained conditions consists of 10 training blocks, followed by two test blocks. Experiment 1 confirmed the findings of Baese-Berk et al. (2013), where training on five unrelated accents facilitated greater comprehension of a new speaker of Mandarin-accented English and a speaker of Russian-accented English. To investigate the effect of lifetime exposure, the current experiment therefore provides no training on foreign-accented speech, instead requiring listeners to first transcribe sentences from native speakers, before being tested on their ability to understand non-native speech. Comprehending speech in noise has been shown to be difficult (e.g. Nábělek, 1988; Nábělek & Dagenais, 1986; Parikh & Loizou, 2005; Pickett, 1957). By providing training on this aspect of the task, but not on L2-accented speech, we ensure participants in this experiment are not at a disadvantage when disambiguating speech in noise.

As in Experiment 1, model talkers for this experiment are from Northwestern University’s Wildcat Corpus of Native- and Foreign-Accented Speech. For the training portion of the attunement task, five male, native speakers of American English were selected who were homogeneous in dialect, as judged by the author and two other trained phoneticians. The setup of this experiment was identical to Experiment 1. The same Mandarin- and Russian-accented model talkers were used at test, and the same sentences were also used, with talker voice and sentence order randomized for each participant.
All participants except those in the no-training condition heard one list of Wildcat sentences during the first set of five training blocks, with voices and sentence order randomized in each block. The second set of five training blocks presented a second list of 15 sentences, also randomized by voice and sentence order in each block. Each of the two test blocks consisted of a new list of 15 sentences, with both Mandarin-accented and Russian-accented recordings, in randomized order. The task during both training and test required participants to orthographically transcribe each sentence. No feedback was provided. Participants heard each sentence only once and began transcribing each one as soon as it started playing. Upon completion of one sentence’s transcription, participants pressed the “Enter” key to hear the next sentence.

3.1.2.2 IAT

Participants completed the IAT following the attunement task, as in Experiment 1.

3.1.2.3 Social network questionnaire

The social network questionnaire was administered following the attunement task (questions are provided in the Appendix and in Chapter 2). Anonymized responses to the questionnaire were reviewed without indicators of recruitment location (NYC or upstate New York) to reduce potential experimenter bias when grouping participants as high- versus low-exposure. Participants were preliminarily categorized as high-exposure if they had more than one family member with whom they interacted regularly and if they answered “yes” to two of the three 30% community questions (i.e. were 30% or more of the students in your elementary school/high school, or 30% or more of the people in your community non-native speakers?).
3.1.2.4 PPVT

Following completion of the social network questionnaire, participants completed the Peabody Picture Vocabulary Test-IV. The PPVT is included to evaluate the effect of an individual’s measured vocabulary size on their word recognition accuracy, as well as to check for differences between participants from different colleges, as a possible explanation for any differences found in performance on the test task.

To test for differences in receptive vocabulary size between the two colleges used for recruitment, linear models of PPVT score by school were implemented in R. Regressions were also performed for proportion of keywords correctly identified in test trials across all participants with fixed effects of PPVT score, condition, and recruitment location.

3.1.2.5 Statistical modeling

Models were constructed in the same manner as in Experiment 1. A base model was constructed of accuracy on keywords with a random effect of participant, then compared to a model with a fixed effect of condition added. Models were compared in a step-up fashion using the \texttt{anova} function until a maximally complex model was constructed that would successfully converge. The resulting logistic mixed effects model of accuracy had fixed effects of training condition, IAT $D$-score, and test language, with random effects of listener and sentence. Recruitment location, keyword position, keyword predictability, and random slopes for test language by participant did not significantly improve model fit and so were excluded from the final model. The reference level for condition was releveled such that both the high exposure condition and the low exposure condition alternately served as reference in the full model.
Models were unable to converge when answers to questions on the Social Network Questionnaire were added. For this reason, a logistic mixed effects regression model was implemented separately to assess the effect of each social network exposure type on accuracy. Accuracy on test trials was entered as the dependent variable and the following were entered as fixed effects: number of non-native accents exposed to, number of “yes” responses for questions (#6-8), number of non-native family members, number of non-native friends, number of non-native other acquaintances (sum of roommates, neighbors, teaching assistants and professors), number of non-native childhood friends (sum of high school and elementary school friends), and number of non-native people participant interacted with on a daily basis. Random effects were included for participant and sentence.

As in Experiment 1, PPVT was assessed via a linear model of proportion accuracy by PPVT score, with additional fixed effects of recruitment location and training condition.

3.2 Results

3.2.1 Social network questionnaire

A logistic mixed effects model of accuracy by responses to the social network questionnaire, with random effects of sentence and participant, identified the number of “yes” responses to the 30% exposure questions (#5-7 on the questionnaire) ($\beta=0.36$, $z=2.49$, $p=0.013$), family ($\beta=0.09$, $z=2.26$, $p=0.024$) and friends ($\beta=0.10$, $z=2.50$, $p=0.012$) as significant predictors. This model was run over all participants. Post-hoc review of participants’ questionnaire responses indicated those preliminarily binned as high-exposure had on average 4.84 family members who were non-native speakers, while those in the low-exposure condition had 0.29 non-native family members, on average. Similarly, high-exposure participants had an average of 6.58 friends who
were non-native speakers, while low-exposure participants had 2.1 friends, on average. The preliminary criteria used for separating participants into condition groups were therefore judged to be sufficient and consistent with the regression analyses.

Untrained participants’ social network questionnaires were evaluated at the time the experiment was administered for each participant to ensure they met the criteria outlined above for categorization as a low-exposure participant.

### 3.2.2 Attunement task

The maximal logistic mixed effects model of accuracy revealed the high-exposure condition ($\mu=98.26\%, \sigma=0.017$) was significantly more accurate than the low-exposure group ($\mu=93.87\%, \sigma=0.034; \beta=1.27, z=6.65, p<0.001$) and the no training group ($\mu=94.7\%, \sigma=0.028, \beta=1.68, z=5.92, p<0.001$) on test trials. Figure 8 plots proportion accuracy on test trials by condition. Test accent (i.e. Mandarin or Russian) was not found to be a significant predictor of accuracy ($\beta=-0.67, z=-1.06, p=0.291$), however a significant effect was found for IAT $D$-score ($\beta=0.9, z=3.66, p<0.001$). $D$-score as a predictor of accuracy is discussed further in section 3.2.3.
As in Experiment 1, recruitment location, keyword position, and keyword predictability did not significantly improve model fit. Whether low-exposure participants received training (no training condition versus training on American English) was not found to be a significant predictor of accuracy (reference level = low exposure; $\beta=-0.41$, $z=-1.57$, $p=0.116$). The average accuracy for both trained low-exposure participants and untrained low-exposure participants was 94%.

### 3.2.3 IAT

Across all participants, those with more positive D-scores were more accurate on keywords ($\beta=0.9$, $z=3.66$, $p<0.001$). This was true of the low-exposure condition ($\beta=0.94$, $z=3.36$, $p<0.001$), however no significant effect of IAT score was found for the high-exposure condition ($\beta=0.35$, $z=0.78$, $p=0.436$).

Different patterns were also found for the two recruitment locations: greater accuracy correlated with greater IAT score among the upstate participants overall (i.e. both high- and low-exposure groups; $\beta=0.91$, $z=2.99$, $p=0.003$), but no such effect was found for the NYC students.

**Figure 8. Proportion accuracy on test trials by exposure and training conditions.**
$(\beta=0.34, z=0.65, p=0.514)$. Among the upstate participants, no effect was found for the high-exposure participants $(\beta=0.57, z=1.45, p=0.148)$, but an effect was found for the low-exposure participants $\beta=0.80, z=3.31, p<0.001$, indicating these participants were driving the effect for the upstate group as a whole. Of note with these analyses, however, is that the high-exposure group primarily consisted of NYC students (24/31), while the low-exposure group was entirely upstate students. The low-exposure upstate participants therefore also appear to be driving the significant effect of IAT score found over all participants.

Figure 9 plots accuracy on test trials by IAT $D$-score. The curved line shows that accuracy increases as IAT score gets increasingly positive. Positive $D$-scores indicate a stronger implicit bias for categorizing non-native speech with the label ‘positive’. Figure 10 provides a scatterplot of IAT scores by accuracy to give a clearer picture of individual participants’ performance.
Figure 9. Plot of accuracy on test trials by IAT score for participants in Experiment 2. Black dots at 0.00 and 1.00 represent each participant’s accuracy for each keyword. Positive $D$-scores indicate a stronger implicit bias for categorizing non-native speech with the label ‘positive’.
3.2.4 PPVT

PPVT scores for each participant are included in the Appendix. A linear model of PPVT score by condition revealed no significant differences between high and low exposure participants in receptive vocabulary score ($\beta=-3.818$, $t=-1.67$, $p=0.101$), nor was participants’ school found to significantly predict their score ($\beta=-2.713$, $t=-1.43$, $p=0.157$).

Receptive vocabulary size was not found to be meaningful in any of the analyses run, neither as a predictor of proportion accuracy on test trials over all participants ($\beta=0.0004$, $t=0.79$, $p=0.435$), nor within separate models for each condition (high: $\beta=-12.83$, $t=-0.199$, $p=0.844$; low: $\beta=-0.0004$, $t=-0.615$, $p=0.543$).
3.3 Discussion

Participants in this study who had more lifetime experience listening to non-native speech, specifically from childhood and from interacting regularly with multiple friends and family members, were better able to accurately identify keywords than less experienced listeners. This finding provides further evidence that prior exposure to numerous non-native accents facilitates an individual’s ability to understand new L2-accented speakers. It shows the results obtained in laboratory attunement studies (e.g. Baese-Berk et al. 2013) are mirrored in the wider population, where training is obtained in more naturalistic settings.

Vocabulary size has been linked to improved comprehension of speech in noise (e.g. Banks et al., 2015; Bent et al., 2016; Tamati et al., 2013). In this experiment, however, no effects were found either within or across participant groups for PPVT score. Students from the upstate college were less accurate than those from the college in NYC, however school was not a significant predictor of PPVT score, suggesting the difference in performance between the two colleges is unlikely due to one group having an unfair advantage on the task due to larger receptive vocabulary abilities.

Analyses comparing those who received no training on speech in noise to participants trained on American English in noise found no evidence that providing participants with such training facilitated recognition of non-native speech in noise. Trained and untrained low-exposure participant groups performed similarly, both achieving a high level of accuracy, but were significantly less accurate than those with greater experience hearing non-native speech over the course of their lifetime. High accuracy rates are discussed further in Chapter 5.
Replicating the difference found here in accuracy between high-exposure participants and low-exposure participants with lower overall accuracy would provide more evidence that increased exposure to non-native accented speech facilitates individuals’ ability to understand new non-native speakers.

3.4 Conclusion

This study investigated the effect of individuals’ lifetime experience listening to non-native accented English on their comprehension of unfamiliar L2 accents and speakers. The results provide evidence in support of the hypothesis that increased experience with a variety of non-native accents in the real world facilitates individuals’ ability to understand non-native speakers. Having further established thus far that exposure both within and beyond the laboratory on non-native speech facilitates attunement and generalization, the next chapter examines what might be causing these effects.
The Category Loosening Hypothesis suggests that when people are faced with accented speech, they shift category boundaries to encapsulate the incoming variation, without reorganizing internal category structure. Previous work has demonstrated that training listeners on multiple speakers of a single accent does not facilitate comprehension of an untrained accent, instead only generalizing to a new speaker of the same accent heard in training (Bradlow & Bent, 2008). The Category Loosening explanation for this would be that because talkers with the same accent are likely to be similar in their pronunciations, due to transference from their L1, the variability of another talker with a different language background is unlikely to be similar enough to fall within listeners’ expanded categories.

However, if listeners are exposed to talkers of several different accents, perhaps their category boundaries would be expanded enough to allow for generalization to a novel L2-accented talker, whose productions might fit within the bounds of what the listener has already heard. If this account is accurate, we might expect greater exposure to native, regionally-accented speakers of English to have similar facilitating effects on comprehension of non-native speech. Though native varieties of English have been found to differ predominantly in vowels (e.g. Adank et al., 2009; Clopper et al., 2005; Nathan et al., 1998), differences are also well-reported in other dimensions, including prosody (e.g. Childs & Wolfram, 2004; Grabe & Post, 2002; Thomas & Carter, 2006; Wells, 1982) and allophonic inventories (e.g. Beal, 2004; Cox & Palethorpe, 2007; Wells, 1982). They have not, however, been shown to be substantially more variable across category types in the way L2 speakers have (Flege et al., 1995; Laturnus, Submitted; Rogers et al., 2012). Training
on different native accents would therefore expose participants to multiple ways of implementing each category, differing substantially from American English categories, without the variability characteristic of non-native speech. This variability between (but not within) speakers will be demonstrated in section 4.1, below.

Alternatively, the nature of the accents may play a crucial role in generalization. If hearing several L1 accents does not generalize to a non-native speaker, it may be that listeners require exposure to L2 speech specifically. The weak interpretation of the Systematic Variability Hypothesis suggests that learning which categories are most commonly difficult for L2 speakers might facilitate comprehension of unfamiliar non-native voices. Features common to non-native English speech include unreduced unstressed vowels (Baker et al., 2011), a lack of the tense/lax contrast (Flege et al., 1999; Tsukada, 2009), and difficulty with certain consonants (e.g. /ɻ/ for Japanese speakers, /θ/ for French speakers), though some of these features may also be found among native varieties (e.g. lack of interdental fricatives in many English dialects, Wells, 1982). The most notable feature, however, is the greater variability of non-native speech relative to native speakers’ productions – though two speakers may deviate from native speakers in the ways listed, they are still likely to differ from one another in their acoustic implementations of any given category. It may be, therefore, that given enough exposure, listeners learn enough about how L2 speech commonly differs from native English that they’re better able to understand even untrained non-native voices.

This question is investigated in the current experiment by training listeners on five, regionally-accented, native speakers of English and testing them on non-native speakers. Is the intra-speaker variation that’s characteristic of L2 speech crucial for generalization to novel non-
native voices? Or do listeners merely need enough exposure to accented speech more generally to improve their comprehension? The Category Loosening Hypothesis predicts that as listeners attune to the native regional varieties heard during training, category boundaries will shift to encapsulate incoming variation, and the variation heard from the L2 test talkers will fit within these expanded categories, facilitating generalization. The former account, however, predicts no generalization effects will be observed in the present study because listeners will not have the requisite exposure to L2 speech to adequately prepare them for the non-native test talkers. Simply hearing different ways categories can be implemented may not be enough to facilitate comprehension of a highly variable L2 accent.

4.1 Experiment 3.1 – Acoustic analysis of native varieties

4.1.1 Methods

Five native varieties of English were chosen that differed from Mainstream American English. One male, native speaker was recorded for each of Jamaican, Manchester, Scottish, New Zealand, and Indian English. These accents were chosen in particular because they are less likely to be as familiar to American listeners as, for example, Received Pronunciation or Australian English (Evans, 2005). The Jamaican English speaker was recorded in a sound-attenuated booth using a head-mounted microphone by the author in New York City. All other speakers were recorded using comparable methods in their country of origin.

Acoustic measurements were taken to correspond to those reported in (Laturnus, Submitted), which were the same non-native voices used during training and test in Chapter 2. These consisted of between-speaker comparisons of voice onset time (VOT), stressed vowel quality and duration, and unstressed vowel quality and duration. Because the same scripted
sentences were used in all experiments, it was possible to extract tokens from the same environments in the current study. Of key interest is the difference in variance between native and non-native English speakers. To this end, the native speakers of American English analyzed in (Laturnus, Submitted) were added to the dataset of regionally-accented native English speakers for all acoustic comparisons.

4.1.1.1 VOT

VOT was measured following the methods and environments used for analysis of the same sentences in Laturnus (Submitted). Measurements were made beginning with the onset of burst energy (rapid change in amplitude) indicating a stop release and ending at the upward zero crossing at the start of periodicity. Both voiced and voiceless stops were measured in word-initial position, at the onset of a stressed syllable, with the aim of extracting 15 tokens for each place of articulation and voicing type, although this was not always possible due to the materials. A script measuring durations and recording the timestamp, filename, word and sentence for each token extracted VOT measures from all 9 speakers. Table 5 provides the number of tokens extracted by place of articulation and voicing, per variety.

| Variety       | Voiceless | | | Voiced | | | |
|---------------|-----------|-----------|-----------|----------|-----------|-----------|
|               | p         | t         | k         | b         | d         | g         |
| Indian        | 12        | 15        | 14        | 13        | 15        | 6         |
| Jamaican      | 12        | 14        | 15        | 13        | 13        | 6         |
| Manchester    | 12        | 14        | 15        | 14        | 14        | 6         |
| New Zealand   | 12        | 14        | 15        | 14        | 14        | 6         |
| Scottish      | 11        | 15        | 15        | 15        | 14        | 6         |
| Total across speakers | 59    | 72        | 74        | 69        | 70        | 30        |

Table 5. Number of tokens extracted by place of articulation and voicing
4.1.2 Vowels

The onset of periodic energy at the start of F2 was taken as the vowel onset, while the offset of this energy in the second formant marked the vowel’s end. For stressed vowel tokens, only those produced in a modal voice that were not pre- or post-approximant and did not occur next to another vowel were included as full-vowel tokens. Formant measures were manually checked prior to extraction via a Praat script. Table 6 provides the total number of vowels of each type that was analysed per speaker. Differences in number of vowels extracted per talker are due to some tokens being too creaky for analysis.

Duration and formant values were measured, hand-checked and extracted from unstressed vowels that occurred adjacent to a stressed syllable.

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>æ</th>
<th>e</th>
<th>ɛ</th>
<th>i</th>
<th>u</th>
<th>o</th>
<th>ə</th>
<th>ʌ</th>
<th>ɔ</th>
</tr>
</thead>
<tbody>
<tr>
<td>IndianEnglish</td>
<td>4</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>15</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Jamaican</td>
<td>3</td>
<td>13</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Manchester</td>
<td>10</td>
<td>13</td>
<td>9</td>
<td>9</td>
<td>15</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>NewZealand</td>
<td>6</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>14</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Scottish</td>
<td>13</td>
<td>9</td>
<td>9</td>
<td>10</td>
<td>13</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 6. Total number of vowel tokens extracted per vowel for each speaker

4.1.2 Results

4.1.2.1 VOT

Levene’s Test of Variance was constructed in R to test for differences in VOT variance between native American English speakers and regionally-accented native speakers. Tests were run separately on voiced and voiceless VOT duration. No significant difference in variance was detected either in voiced VOT (F=0.44, p=0.82) or in voiceless VOT (F=1.7, p=0.14), indicating all native English speakers were similarly consistent in their productions. Significant differences
were found, however, in both voiced (F=2.29, p=0.035) and voiceless (F=2.12, p=0.05) VOT durations when Levene’s Test was run over the collective dataset of native American English speakers and non-native speakers used for training in Chapter 3 and analyzed in Laturnus (Submitted).

Experiment 3 was designed to test whether exposure to substantial between-speaker variation would facilitate generalization. Therefore, subsequent analyses on extracted acoustic measurements were run in order to verify that the regional varieties of English differed in how categories were implemented. Linear models were constructed in R with log-transformed VOT\textsuperscript{8} as the dependent variable and speaker and place (labial, alveolar, velar) as independent variables. The American English speaker was set as the reference level for speaker, while alveolar was the reference level for place. For voiceless tokens, there was a significant effect for Indian English ($\beta = -0.48, t=-4.06, p<0.001$), with shorter VOT durations than American English, and Jamaican English ($\beta = 0.22, t=1.92, p=0.05$), with longer. No significant effects were found for place, though there was a significant interaction between speaker and labial place for Manchester English ($\beta = -0.37, t=-2.22, p=0.03$). Labial position set as the reference level revealed a trending effect for alveolar place ($\beta = 0.20, t=1.71, p=0.09$), and significant interaction effects for Manchester and alveolar place ($\beta = 0.37, t=2.22, p=0.03$), Manchester and velar place ($\beta = 0.37, t=2.12, p=0.03$), and Scottish and velar place ($\beta = 0.45, t=-2.57, p=0.01$). A trending interaction between Indian English and velar place was additionally found ($\beta = 0.32, t=1.89, p=0.06$).

\textsuperscript{8} Because VOT measurements had a skewed distribution, log-transformed values were used in the models (following e.g., Fricke 2013, Baese-Berk & Goldrick 2009, Stuart-Smith et al. 2015, Piccinini & Arvaniti 2015)
A linear model of log VOT by speaker and place for the voiced tokens shows a significant effect for Indian English ($\beta=-0.46$, $t=-3.31$, $p=0.001$), again with having shorter voiced VOT than American English, and for Jamaican ($\beta=-0.93$, $t=-6.44$, $p<0.001$), also shorter. The linear model additionally revealed significant effects for both labial place ($\beta=-0.50$, $t=-3.61$, $p<0.001$), whose durations were shorter than alveolars, and velar place ($\beta=0.567$, $t=3.11$, $p=0.002$), whose durations were longer. Interactions of speaker by place were found for Indian English and velar place ($\beta=-0.53$, $t=-2.07$, $p=0.04$) and Scottish English and velar place ($\beta=-0.63$, $t=-2.46$, $p=0.015$), and a trending effect for New Zealand English and labial place ($\beta=0.35$, $t=1.79$, $p=0.075$). With labial set as the reference level, there was a significant effect of place for both alveolar ($\beta = 0.50$, $t=3.62$, $p<0.001$) and velar ($\beta = 1.07$, $t=5.92$, $p<0.001$). Interactions were also found between speaker and velar place for Indian English ($\beta = -0.83$, $t=-3.2$, $p=0.002$) and Scottish English ($\beta = -0.68$, $t=-2.65$, $p=0.009$), with trending effects for New Zealand and velar place ($\beta = -0.68$, $t=-2.65$, $p=0.06$), and New Zealand and alveolar place ($\beta = -0.35$, $t=1.79$, $p=0.07$).

Average VOT duration and standard deviation is provided in Tables 7 and 8 for regionally-accented speakers and all speakers used for analyses from Laturnus (Submitted). Plots of average VOT duration for each speaker are provided in Figures 11 and 12.
<table>
<thead>
<tr>
<th>Variety</th>
<th>Average Voiced VOT</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>American English 1</td>
<td>24.25</td>
<td>9.50</td>
</tr>
<tr>
<td>American English 3</td>
<td>24.22</td>
<td>9.61</td>
</tr>
<tr>
<td>American English 4</td>
<td>19.37</td>
<td>9.60</td>
</tr>
<tr>
<td>India</td>
<td>15.13</td>
<td>6.02</td>
</tr>
<tr>
<td>Jamaica</td>
<td>12.30</td>
<td>8.48</td>
</tr>
<tr>
<td>Manchester</td>
<td>19.51</td>
<td>8.54</td>
</tr>
<tr>
<td>New Zealand</td>
<td>21.84</td>
<td>9.25</td>
</tr>
<tr>
<td>Scotland</td>
<td>18.88</td>
<td>9.87</td>
</tr>
<tr>
<td>Chinese L2</td>
<td>27.11</td>
<td>21.85</td>
</tr>
<tr>
<td>Farsi L2</td>
<td>17.23</td>
<td>11.31</td>
</tr>
<tr>
<td>Italian L2</td>
<td>16.02</td>
<td>9.94</td>
</tr>
<tr>
<td>Korean L2</td>
<td>20.91</td>
<td>13.75</td>
</tr>
<tr>
<td>Russian L2</td>
<td>20.94</td>
<td>14.66</td>
</tr>
<tr>
<td>Thai L2</td>
<td>22.94</td>
<td>14.30</td>
</tr>
</tbody>
</table>

Table 7. Average VOT duration and standard deviation for voiced tokens by English variety.

<table>
<thead>
<tr>
<th>Variety</th>
<th>Average Voiceless VOT</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>American English 1</td>
<td>58.68</td>
<td>18.31</td>
</tr>
<tr>
<td>American English 3</td>
<td>58.77</td>
<td>20.92</td>
</tr>
<tr>
<td>American English 4</td>
<td>60.11</td>
<td>17.00</td>
</tr>
<tr>
<td>India</td>
<td>40.30</td>
<td>19.34</td>
</tr>
<tr>
<td>Jamaica</td>
<td>66.26</td>
<td>22.92</td>
</tr>
<tr>
<td>Manchester</td>
<td>49.67</td>
<td>15.42</td>
</tr>
<tr>
<td>New Zealand</td>
<td>55.63</td>
<td>17.95</td>
</tr>
<tr>
<td>Scotland</td>
<td>49.15</td>
<td>18.78</td>
</tr>
<tr>
<td>Chinese L2</td>
<td>73.68</td>
<td>25.35</td>
</tr>
<tr>
<td>Farsi L2</td>
<td>96.79</td>
<td>20.37</td>
</tr>
<tr>
<td>Italian L2</td>
<td>46.77</td>
<td>23.32</td>
</tr>
<tr>
<td>Korean L2</td>
<td>94.21</td>
<td>27.81</td>
</tr>
<tr>
<td>Russian L2</td>
<td>41.27</td>
<td>22.12</td>
</tr>
<tr>
<td>Thai L2</td>
<td>82.7</td>
<td>20.45</td>
</tr>
</tbody>
</table>

Table 8. Average VOT duration and standard deviation for voiceless tokens by English variety.
Figure 11. Average VOT duration by English variety. English 1, 3, and 4 refer to the native speakers of American English. Error bars indicate standard error.

4.1.2.2 Stressed vowels

The standard deviation of F1 and F2 was calculated for each stressed vowel per talker as a measure of variance. This is in line with analyses conducted by Wade et al. (2007), who found non-native speakers were 33% more variable than the native speakers in his experiment. Similar analyses of variance in the non-native speakers used in training in Chapter 2 found they were on average twice as variable as the native speakers (Laturnus, Submitted).

Linear models of standard deviation by speaker were implemented in R and run over the regionally-accented native English varieties along with the native American English speakers. To
test whether the native regional varieties were more or less variable than the American English speakers, the first native speaker of American English was set as the baseline. Laturnus (Submitted) found significant effects of speaker in F1 and F2 variance for three of the five non-native speakers analyzed. In the current data, however, no significant effects of speaker were found in either F1 or F2 variance.

To test for global differences over the entire vowel space for each speaker, linear models were constructed of F1 and F2 by variety, with vowel as an additional fixed effect, following (Chang, 2012; Laturnus, Submitted). American English was once again set as the reference level, along with /a/ for vowels. All vowels were significant predictors of F1 except for /æ/ ($\beta=9.88$, $t=7.77$, $p=0.44$), which varied considerably in F1 values between talkers. Because /a/ was set as the baseline, significant effects for vowel indicate these categories differed in F1 from /a/ over all speakers. Changing the baseline of the model changes which vowels show significant effects (e.g. with /e/ as the baseline, no significant effects are seen for /u/, /o/, or /u/). Significant effects of speaker were found for Manchester English ($\beta=27.99$, $t=2.63$, $p=0.009$), New Zealand English ($\beta=-29.02$, $t=-2.63$, $p=0.009$), and Scottish English ($\beta=-84.78$, $t=-7.90$, $p<0.001$), suggesting these speakers differed in their F1 ranges compared to American English.

All vowels except /o/ ($\beta=-59.63$, $t=-1.44$, $p=0.15$), /u/ ($\beta=77.59$, $t=1.93$, $p=0.055$), and /ʌ/ ($\beta=6.57$, $t=0.16$, $p=0.876$) were significant predictors of F2. Significant effects of speaker were found for Manchester ($\beta=-265.68$, $t=-9.93$, $p<0.001$) and Scottish ($\beta=-168.27$, $t=-6.24$, $p<0.001$), along with a trending effect for Indian English ($\beta=-49.55$, $t=-1.79$, $p=0.073$). Figures 12 to 17 provide plots of the vowel space for each non-native variety (purple) alongside the American English speaker (grey), reproduced here from Laturnus (Submitted). Figures 18 to 22 provide
comparable vowel plots for each regional variety (teal) alongside the American English speaker (red), who was used as the baseline in the linear models.

Figure 12. Vowel plot of Chinese-accented English and American English.
Figure 13. Vowel plot of Korean-accented English and American English.

Figure 14. Vowel plot of Thai-accented English and American English.
Figure 15. Vowel plot of Russian-accented English and American English.

Figure 16. Vowel plot of Farsi-accented English and American English.
Figure 17. Vowel plot of Italian-accented English and American English.
Figure 18. Vowel plot of Indian English and American English.
Figure 19. Vowel plot of Jamaican English and American English
Figure 20. Vowel plot of Manchester English and American English.
Figure 21. Vowel plot of New Zealand English and American English.
4.1.2.3 Unstressed vowels

A linear model of unstressed vowel duration by speaker returned no significant effects. This is in contrast to the analyses conducted on non-native data (Laturnus, Submitted), where non-native speakers were an average of 31 ms longer in schwa duration than the native American English speakers. Table 9 shows how the difference is smaller when calculated over all native English speakers, however L2 speakers are still three times more variable in their unstressed vowel durations ($\mu=73$ ms, $SD=33.8$) compared to the native speakers ($\mu=52.17$ ms, $SD=11.25$). This is likely due to non-native speakers’ tendency not to fully reduce all their unstressed vowels (e.g. Busà, 2010; Choi, 2005; Kim & Lee, 2005; Laturnus, Submitted; Mok & Dellwo, 2008).
Linear models of F1 values by variety for schwa revealed significant effects for Jamaican English ($\beta=152.63$, t=3.80, p=0.001) and Indian English ($\beta=85.07$, t=2.01, p=0.05), and a trending effect for Manchester English ($\beta=76.2$, t=1.9, p=0.073). Linear models of F2 for schwa found a significant effect only for the Manchester English speaker ($\beta=-180.65$, t=-0.07, p=0.049).

<table>
<thead>
<tr>
<th>Variety</th>
<th>Mean /ə/ duration</th>
<th>SD of /ə/ duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>American English 1</td>
<td>36.41</td>
<td>9.83</td>
</tr>
<tr>
<td>American English 3</td>
<td>44.99</td>
<td>10.02</td>
</tr>
<tr>
<td>American English 4</td>
<td>42.78</td>
<td>12.31</td>
</tr>
<tr>
<td>India</td>
<td>64.46</td>
<td>11.51</td>
</tr>
<tr>
<td>Jamaica</td>
<td>83.97</td>
<td>12.69</td>
</tr>
<tr>
<td>Manchester</td>
<td>53.49</td>
<td>20.16</td>
</tr>
<tr>
<td>New Zealand</td>
<td>53.49</td>
<td>9.62</td>
</tr>
<tr>
<td>Scotland</td>
<td>37.75</td>
<td>3.85</td>
</tr>
<tr>
<td>Chinese L2</td>
<td>75.35</td>
<td>37.44</td>
</tr>
<tr>
<td>Farsi L2</td>
<td>79.84</td>
<td>61.58</td>
</tr>
<tr>
<td>Italian L2</td>
<td>54.08</td>
<td>14.62</td>
</tr>
<tr>
<td>Korean L2</td>
<td>103.28</td>
<td>48.57</td>
</tr>
<tr>
<td>Russian L2</td>
<td>72.32</td>
<td>13.26</td>
</tr>
<tr>
<td>Thai L2</td>
<td>52.94</td>
<td>27.30</td>
</tr>
</tbody>
</table>

Table 9. Average schwa duration and standard deviation per speaker. Bold line separates native from non-native speakers.

Each native English speaker was consistent in their productions of schwa within the vowel space; linear models of F1 and F2 standard deviations by speaker returned no significant results. Figure 23 provides a plot of all native varieties’ schwa productions.
Figure 23. Unstressed vowel productions for each native English speaker. Ellipses indicate 95% confidence intervals.

4.1.3 Discussion

As expected, acoustic analyses of the regionally-accented native speakers suggest they vary in their category implementations relative both to each other and to American English. They are additionally as consistent in their vowel and VOT productions as American English speakers, and more so than non-native speakers. These five native speakers of different regional varieties were used as model talkers in the training portion of the attunement task in Experiment 2, to investigate whether training on different native accents facilitates generalization to non-native accented speech.
4.2  Experiment 3.2 – Perceptual adaptation to native varieties of English

4.2.1  Methods

4.2.1.1  Participants

Thirty-one monolingual participants with no language-related disabilities or diagnosed speech or hearing disorders participated in this study for $15. The study took an average of 45 minutes to complete.

All participants were freshmen students at a college in New York City who self-reported having limited prior exposure to non-native speech. They were not born or raised in a major metropolitan area and had lived in New York City for less than six months. A social network questionnaire, described in detail in §3.1.5, verified that participants’ prior experience with non-native speech was consistent with other studies’ requirements for low exposure (Experiments 1 and 2).

4.2.1.2  Materials and procedure

4.2.1.2.1  Attunement task

The design of this experiment was identical to that of the two previous studies (Experiments 1 and 2), with the exception of the training voices. During training, participants in this experiment heard and transcribed the five native English speakers who were described and acoustically analyzed above in §2. The same sentences were heard during training as in the previous two studies, with sentence order and talker voice randomized across sets for each participant. The sentences and voices heard during test were identical to the two previous studies, with sentence order and talker voice also randomized for each participant.
Accuracy was measured on the same keywords chosen for each sentence as in Chapters 2 and 3. Performance on test trials was compared to the low-exposure condition and the no-training condition in Chapter 3. Low-exposure participants in that study were trained on five male, native speakers of American English selected from the Wildcat Corpus (Van Engen et al., 2010), while the no-training group only completed the test blocks. Performance was additionally compared to the experimental condition from Chapter 2, who were trained on five non-native English speakers. This allows for an evaluation of the effectiveness of native versus non-native input during training in generalization to novel non-native speakers.

4.2.1.2.2 Implicit Association Test, social network questionnaire, and PPVT

The Implicit Association Test (IAT), social network questionnaire, and Peabody Picture Vocabulary Test (PPVT) administered in this experiment were identical to that described in the previous two chapters.

4.2.1.2.3 Statistical modeling

Models were constructed in the same manner as in Experiments 1 and 2. A base model was constructed of accuracy on keywords with a random effect of participant, then compared to a model with a fixed effect of condition added. Models were compared in a step-up fashion using the anova function until a maximally complex model was constructed that would successfully converge. The resulting logistic mixed effects model of accuracy had fixed effects of training condition, IAT D-score, and test language, with random intercepts for listener and sentence and random slopes for test language by participant. Recruitment location, keyword position, and keyword predictability did not significantly improve model fit and so were excluded from the final model. The reference level for condition was relevelled between runs of the full model so that full
comparisons could be made between participants trained on native English varieties, American English, and L2-accented English.

PPVT was assessed via a linear model of proportion accuracy by PPVT score, with additional fixed effects of recruitment location and training condition.

4.2.2 Results

4.2.2.1 Social network questionnaire

Table 10 is adapted from Table 3 in Chapter 2, with the average responses to each question in the social network questionnaire from the current experimental group added. A mixed-effects logistic regression model of accuracy by each question in the questionnaire found no significant predictors. This is interpreted as being due to the homogeneity of the participant pool, all of whom met the requirements identified in Chapter 3 for low exposure: one or fewer non-native family member with whom they interacted regularly, and responding “yes” to no more than one of the three 30% community questions (ne-6 on the questionnaire: at least 30% of high-school, elementary, or general community growing up were non-native English speakers).
<table>
<thead>
<tr>
<th>SNQ question (number of)</th>
<th>L2-trained group average (Experiment 1)</th>
<th>Control group average (Experiment 2)</th>
<th>Regional variety trained average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-native accents exposed to</td>
<td>2.86</td>
<td>2.91</td>
<td>3.23</td>
</tr>
<tr>
<td>Family members</td>
<td>0.3</td>
<td>0.29</td>
<td>0.03</td>
</tr>
<tr>
<td>Friends</td>
<td>1.54</td>
<td>2.1</td>
<td>2.3</td>
</tr>
<tr>
<td>Other (profs, TAs, neighbours, roommates)</td>
<td>4.46</td>
<td>4.7</td>
<td>5.97</td>
</tr>
<tr>
<td>Childhood friends (high school, elementary school)</td>
<td>1.6</td>
<td>2.7</td>
<td>3.73</td>
</tr>
<tr>
<td>Non-native accents with daily exposure</td>
<td>1.29</td>
<td>2.12</td>
<td>2.5</td>
</tr>
<tr>
<td>30% community (#4-6)</td>
<td>0.21</td>
<td>0.23</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Table 10. Average number of people exposed to in each category of the social network questionnaire by training condition.

4.2.2.2 Attunement task

A logistic mixed effects regression model of accuracy by training condition found participants trained on non-native varieties of English were significantly more accurate on test trials than those trained on native English varieties ($\beta=0.36$, $z=2.04$, $p=0.042$). No significant effects were found for the American English-trained participants compared to the regionally-trained participants ($\beta=-0.12$, $z=-0.69$, $p=0.492$). Figure 24 plots proportion accuracy for keywords by condition.
Figure 24. Proportion accuracy on test trials by training group. “Native” refers to the current study’s experimental group, who were trained on native regional varieties of English.

Figure 25 shows how participants improved their accuracy over the course of the experiment. This is consistent with non-native training data (Experiment 1), but not with American English training data, where participants’ accuracy was already at ceiling upon first exposure (Chapter 3).
Fixed effects of recruitment location ($\chi^2(2)=2.34$, $p=0.311$), position ($\chi^2(2)=0.74$, $p=0.692$), and predictability ($\chi^2(1)=2.63$, $p=0.100$) did not significantly improve model fit. As in the previous two experiments, test accent (i.e. Mandarin or Russian) did improve model fit ($\chi^2(1)=23.51$, $p<0.001$), however performance on the two test voices did not significantly differ over all participants ($\beta=-0.62$, $z=-1.00$, $p=0.317$).
4.2.2.3 Implicit Association Test

Across regionally-trained participants, those with more positive D-scores were more accurate on key words ($\beta=0.03$, $z=3.34$, $p=0.001$), which is consistent with past results (Experiment 1). An additional logistic mixed-effects model with a random effect of participant was run over the aggregate date from all 143 participants in this dissertation. It revealed that IAT score significantly predicted accuracy ($\beta=-0.53$, $z=2.82$, $p=0.005$), which is unsurprising given that this effect has been observed in every experiment reported here with the exception of the high-exposure lifetime group in Experiment 2, who did not show a significant effect.
Figure 26. Plot of accuracy on test trials by IAT score for participants in Experiment 3. Black dots at 0.00 and 1.00 represent each participant’s accuracy for each keyword. Positive D-scores indicate a stronger implicit bias for categorizing non-native speech with the label ‘positive’.
Figure 27. IAT score by proportion accuracy for participants in Experiment 3. A more negative IAT score indicates a stronger bias for associating non-native English with the label “negative”, while a more positive IAT score indicates the opposite bias.

4.2.2.4 PPVT

PPVT scores for each participant are included in the Appendix. Experiment group was not a significant predictor of PPVT score (native regional training was set as the reference level; American English training: \( \beta = -2.9, t = -1.276, p = 0.205 \); non-native training \( \beta = -0.196, t = -0.09, p = 0.926 \)). PPVT score was not a significant predictor of accuracy within the group trained on native varieties (\( \beta = 0.0007, t = 1.053, p = 0.301 \)), however an effect was found when the model was run over all collected data (i.e. all participants from Chapters 2 through 4; \( \beta = 0.0008, t = 2.259, p = 0.026 \)).
4.2.3 Discussion

In the attunement portion of the experiment, no significant effect of training group was found for those trained on native speakers of American English or those trained on native regional varieties. The group trained on non-native varieties of English (from Chapter 2) significantly outperformed both of these conditions.

The lack of improved accuracy on the Mandarin- and Russian-accented trials compared to the control group suggests native and non-native speech are processed differently. Experience hearing multiple examples of how different varieties of English can differ from the listener’s own speech does not appear to be adequate for generalization to previously unheard non-native voices. The Weak Systematic Variability Hypothesis would account for this finding by highlighting the importance of training on various L2 accents in order to learn the similarities found across different talkers. While native regional varieties of English do differ from Mainstream American English and from each other in many ways, they do not have the same characteristic variability or lack of key contrasts that are common to many non-native accents. Training on L1 varieties therefore does not provide listeners with the benefit of experience mapping between this greater intra-speaker variability and phonological categories or lexical items, which results in lower accuracy. The lack of generalization to the non-native speakers also suggests the Category Loosening Hypothesis may not be an accurate account of the mechanism underlying adaptation and generalization. The native model talkers did differ from each other in how they produced their phonological categories. If listeners had expanded category boundaries to encapsulate this variability, their resulting categories likely would have at least partially overlapped those of the non-native test talkers, facilitating faster adaptation.
Past experiments in the current body of work have not found PPVT score to significantly predict accuracy, however with the data from all three experiments combined, an effect does emerge. This effect is consistent with prior literature, which has connected greater receptive vocabulary scores on the PPVT to greater accuracy in word recognition for speech in noise (e.g. Banks, Gowen, Munro, & Adank, 2015; Bent, Baese-Berk, Borrie, & McKee, 2016; Tamati, Gilbert, & Pisoni, 2013). Because no significant effect of experiment group was found, as in previous chapters, this suggests that while some participants may have benefitted from greater receptive vocabulary abilities, they do not appear to have been unequally distributed between conditions. Differences found between experimental groups in the speech transcription task, therefore, are unlikely due to the receptive vocabulary sizes of participant groups.

Participants in this study who had stronger implicit biases against non-native speech, as measured by the IAT, performed significantly less accurately on test trials than those with less stereotypical biases. This is consistent with the results reported in Chapter 2.
Chapter 5: General discussion & further directions

In the next section, the Ideal Adapter Framework is applied to the results from the current body of work as a means of explaining how adaptation and generalization might have proceeded for participants in each experiment. Following this, a summary and general discussion is offered for each of the Implicit Association Test and Peabody Picture Vocabulary Test that were conducted in each experiment, followed by a discussion of the limits in the current methodology that should be taken into consideration in future work.

5.1 The Ideal Adapter Framework and current data

Experiment 1 was designed to replicate the results of Baese-Berk et al. (2013) while additionally investigating the effect of implicit bias on attunement. Participants who received training on five unrelated non-native accents had greater accuracy on test trials than the control group, who were trained instead on five native speakers of American English.

Based on the explanation of the framework outlined by Kleinschmidt and Jaeger (2015), discussed in Chapter 1, the proposal for this experiment is that generalizing across different non-native accents would require constructing a model of L2 speech based on the similarities in how different talkers’ speech deviates from native English (e.g. greater variability, unreduced unstressed vowels, or the lack of a tense/lax contrast). The suggestion is that these patterns need not constitute close acoustic category matches between speakers, mainly because little close acoustic overlap was observed between the different speakers used in Experiment 1 (Laturnus, Submitted). Additionally, non-native accents are often marked by transference effects from their native language (Flege et al., 1995), along with inconsistency in pronunciations (Rogers et al.,
Constructing the model thus doesn’t necessitate that no acoustic information be tracked and incorporated, but instead that in addition to what acoustic similarities are present between talkers, the model include more general patterns that might speed adaptation to novel L2-accented talkers. Examples of possible patterns include greater within-speaker variation, a lack of certain segments or contrasts (e.g. interdental fricatives, tense/lax vowels, differentiating /l/ and /ɹ/), or atypical stress patterns, as discussed in Chapter 1. Building such a model would necessitate the listener having enough exposure to non-native speech to recognize these commonalities. They would also need a reasonable expectation, based on past experience, that grouping these talkers together would facilitate comprehension of new, similar voices. Participants in Experiment 1, along with those of Baese-Berk et al. (2013), would begin by constructing talker-specific models for each of the voices they heard and would construct a more general, “moderately-accented, male non-native speaker of English” model as the experiment progressed. This model would then be applied to the Mandarin and Russian speakers at test, resulting in higher accuracy than those who were not trained on the non-native speakers, and thus lacked a unifying model. Because non-native speakers are inconsistent in their pronunciations, listeners with relatively little experience with L2 speech are likely to apply previous models to new L2-accented speech with a greater degree of uncertainty than if they were native speakers. The greater inconsistency in production found among non-native talkers is likely to require constant learning on the part of the listener. Due to lack of input, inexperienced listeners might also choose the wrong model to apply to an unfamiliar L2-accented voice. This would necessitate more input and adaptation, resulting in slowed recognition.

Experiment 2 extended the findings of Experiment 1 by comparing populations who differed in their level of real-world exposure to non-native speech beyond the laboratory. The goal
of this was to investigate whether the pattern found in studies like the one presented in Chapter 2, replicating the findings of Baese-Berk et al. (2013), is mirrored in the wider population, who must attune to non-native speech and generalize to new talkers in more natural settings than in the laboratory. Participants who reported higher overall exposure to non-native speech, particularly from family, friends, and during childhood, were more accurate in transcribing two previously unheard speakers than were those with limited prior exposure to non-native speech. The same mechanism proposed to account for the findings of Experiment 1 is applicable to Experiment 2, with a crucial difference being that the high-exposure participants in Experiment 2 would have built up their models of non-native speech over the course of their lifetime through their interactions with L2 speakers. This is in contrast to the participants in Experiment 1, who would have built their models merely during the approximately 35-minute training that was provided during the experiment, since they had limited prior exposure to L2 speech before entering the lab. These models may also be more fragile for the less experienced listeners, based only on brief input acquired in an artificial setting, and so might not extend as robustly or last as long outside the laboratory as those who built their models over a longer period of time through natural interactions with speakers. The finding that the high-exposure participants in Experiment 2 significantly outperformed the participants trained on non-native speech in Experiment 1 is consistent with this explanation, since the former’s models would have had substantially more input from a variety of non-native accents and therefore should be more adept at generalizing to novel non-native talkers and be applied with a greater level of certainty.

The results from Experiments 1 and 2 demonstrate that there is a benefit of non-native training, whether in-lab or over the lifetime, for attunement and generalization to unfamiliar non-
native talkers. These studies, however, are not able to speak directly to the plausibility of the hypotheses described in Chapter 1. Experiment 3, in conjunction with Laturnus (Submitted), was designed to shed further light on whether the Category Loosening Hypothesis or Weak Systematic Variability Hypothesis might be better able to account for the effects reported here and in the literature. Exposure to native varieties of English provided listeners with examples of between-talker variability. The suggestion of the former account was that listeners only need examples of between-talker variability to expand their categories enough such that new incoming speech is likely to fall within the bounds of those categories. The second proposal was that the within-speaker variability common to L2 speech would be instrumental for generalization to novel L2 speakers, being a defining feature that distinguishes it from native-accented speech. The results of this experiment are consistent with the latter hypothesis and are more fully explained within the interpretation of the Ideal Adapter Framework presented here. While native regional varieties of English do differ from Mainstream American English and from each other in many ways, there are also key ways they differ from non-native speech that might prevent them from being incorporated into the same model as L2 speakers. Chief among them are a lower likelihood of hesitancies, stuttering or disfluency (Cucchiarini, Strik, & Boves, 2000; Möhle, 1984); a lower likelihood of collapsing the tense/lax contrast common to most native dialects (Pennington, 2014); and lower overall variability in production compared to non-native speakers (Chapter 4; Laturnus, Submitted). In Experiment 3 (Chapter 4), then, participants would have built talker-specific models for each voice they heard in training. Whether a more general model would be constructed for “male, native speaker of regionally-accented English” from this training data is an open question not tested here, though it would be predicted if commonalities exist between different native
regional dialects. When faced with the test blocks, participants would likely recognize that the incoming speech did not match any of their already existing models and so would engage in active statistical learning, quickly attuning to these new speakers (Clarke & Garrett, 2004), but nevertheless achieving worse accuracy on the new non-native test speakers than those trained on non-native speakers (Chapter 2) or having substantial lifetime experience hearing non-native speech (Chapter 3).

5.2 Implicit Association Test

An Implicit Association Test measured participants’ implicit biases towards non-native accented speech by requiring them to quickly categorize speech and orthographic words as either ‘accented’/‘unaccented’ or ‘positive’/‘negative’, respectfully. The co-occurrence of both sets of labels on the screen during each test trial made the task of ignoring irrelevant labels difficult, thereby measuring participants’ readiness to categorize, for example, non-native accented speech as ‘negative’. Scores on the IAT were used to predict individuals’ accuracy in transcribing the Mandarin- and Russian-accented voices during the attunement task. The overall pattern that emerged from these tests was that those with stronger biases for categorizing non-native accented speech as ‘negative’ had less accuracy on test trials during the attunement task. One explanation for this results is that participants who view non-native speech more negatively might have a greater expectation for unintelligibility or incoherence, and so exert less effort in comprehension (Kang & Rubin, 2009; Lippi-Green, 1997; Rubin, 1992).

Several studies have demonstrated that being led to believe a talker is of Asian descent can affect listeners’ perception in audiovisual tasks (e.g. Babel & Russell, 2015; Fraser & Kelly, 2012; Kang & Rubin, 2009; Rubin, 1992), potentially due to a greater expectation for hearing L2-
accented speech, even when the model talker speaks English natively (see summary of the literature in Chapters 1 and 2). The experiments presented in this body of work found more negative biases towards non-native English did correlate with worse comprehension on the transcription task. If this is because participants have a lower expectation for fluency, comprehensibility or coherence, it could be that they exert less effort in mapping between variable input and lexical categories. In hearing a native voice, for example, utter a sentence that sounds like “when she raised in a few they eat grass”, listeners might take the extra step of inferring the speaker probably meant something more like “when sheep graze in a field they eat grass”, since that is a more comprehensible sentence. Upon hearing the same sentence in a non-native voice, however, listeners might not make any such inferences if they have a lower expectation of fluency, and so make less of an effort to interpret nonsensical utterances.

This pattern of more negative bias correlating with worse performance was consistent between the three experiments and across all participants when regressions were run over the collective dataset. The only experimental condition that did not exhibit this pattern was the high-exposure group in Experiment 2, who still exhibited a similar range of both negative and positive IAT scores. One potential explanation for this is that the high-exposure participants were already at ceiling in their accuracy, where it is difficult to measure significant effects. Alternatively, participants’ models of L2 accuracy could be adequately well-developed to override implicit bias, eventually learning the patterns common to different L2 talkers despite being less consistent or efficient in their mappings between variability and lexical categories. Because they would have had more exposure to non-native speech, they would be able to apply their model of non-native speech to novel talkers with a higher degree of certainty than participants whose models were
developed from more limited input, as would be the case with participants trained on L2 speech in Experiment 1. This is a more comprehensive explanation of the pattern, since Experiment 1 participants were also near ceiling, but still displayed the same pattern of worse accuracy for those with more negative biases (accent-trained condition: 96%, high-exposure condition: 98%).

It is also possible participants perceive non-native accented speech negatively without expecting it to be incomprehensible. Many of the high-exposure participants come from families with several non-native speakers. Research in sociolinguistics has demonstrated that immigrant communities often have negative views of their own speech, often even more so than less recent immigrants (i.e. monolingual English-speaking Americans) (e.g. Dailey, Giles, & Jansma, 2005; Derwing, 2003; Fayer & Krasinski, 1987; Riches & Foddy, 1989). These high-exposure participants may therefore have negative views of non-native speech without having greater expectations for incomprehensibility when encountering unfamiliar L2 speakers. Substantial past experience with L2-accented speech would facilitate their comprehension of other non-native accents despite persistent negative biases towards the non-native accented nature of the speech.

5.3 Role of receptive vocabulary

The Peabody Picture Vocabulary Test was administered to all participants in these experiments to measure their receptive vocabulary abilities. Receptive vocabulary size has been linked to greater accuracy for speech comprehension in noise (Banks et al., 2015; Bent et al., 2016; Tamati et al., 2013). No significant differences in vocabulary size were found in any of the experiments between individual conditions or recruiting locations. PPVT was found to be a significant predictor of accuracy within the group of participants trained on non-native speech in Experiment 1. It was also found to be a significant predictor of accuracy when regressions were
run over the data for all 143 participants, with higher vocabulary scores predicting higher accuracy on test trials. The lack of effects found between participant groups suggests participants were not unequally divided between conditions such that any one condition had an advantage in comprehension due to greater vocabulary size. The lack of effects within conditions, other than that found for the accent-training condition in Experiment 1, may be due to a combination of methodological differences between the current study and those that report significant effects of receptive vocabulary score on word recognition in noise. For example, Banks et al. (2015) tested participants on the Harvard Sentences (IEEE, 1969), which are longer, on average, than the sentences used in the current work, as well as being phonetically balanced and consistently low in predictability. Banks additionally used an adaptive staircase procedure, specifically designed to minimize ceiling effects resulting from perceptual adaptation: as participants adapted to the accented talker, the SNR decreased, making the task more difficult. The participants in Tamati et al.’s (2013) study heard Perceptually Robust English Sentence Test Open-set (PRESTO) sentences (Gilbert, Tamati, & Pisoni, 2013), which consist of varying syntactic structures and lexical items of varying difficulty. The purpose of the PRESTO sentences is to increase listening demands by decreasing the predictability of the material listeners hear. These sentences were also presented in eight different dialects, embedded in multitalker babble, to the participants in Tamati et al. (2013) and Gilbert et al. (2013). The current experiment may therefore be easier for participants, using more predictable sentences, consistent SNRs that allowed for adaptation, and using a different kind of background noise than the multitalker babble used by Tamati et al. (2013).
5.4 Shortcomings of the current methodology and considerations for future work

Overall accuracy rates in all three experiments were very high, particularly compared to previous research using comparable methods. Average accuracy for control participants on test trials in Bradlow and Bent (2008), was approximately 75%. In Experiment 1, participants who received training on non-native accented speech were significantly more accurate than the control group, however even the control group had an average accuracy of 94%. High-exposure participants in the lifetime study (Experiment 2) had an average accuracy of 98%. Native listeners rated the talkers used here, which were drawn from the Wildcat corpus, for accentedness on a 9-point scale, while those in NUFAESD were rated for intelligibility. Research comparing listener’s ratings of accentedness relative to perceived comprehension and actual intelligibility, however, suggest this is a difficult task that people aren’t very good at. As mentioned in Chapter 1, (Munro & Derwing, 1995) found their participants rated non-native speakers as strongly accented and difficult to understand even when they were able to transcribe the talker’s speech with perfect accuracy. They additionally observed that accent ratings varied considerably in general in the experiment, regardless of actual intelligibility, with the majority of scores skewed towards more accented. It is therefore likely the case that by choosing L2 speakers with mid-range scores for accentedness, the talkers were actually more intelligible than those in Baese-Berk et al. (2013), leading to higher accuracy for participants, despite having the same SNR.

Given the effect of lifetime exposure found in the data presented in Chapter 3, alongside the improvement following laboratory training reported in prior work (e.g. Baese-Berk et al. 2013), another possibility for the high accuracy in these studies is that the control participants in previous work had less lifetime experience with non-native speech. Despite screening individuals for
exposure to foreign-accented speech and actively recruiting inexperienced listeners in all but Experiment 2’s high-exposure participant group, nearly all of the untrained participants in this dissertation reported contact with L2 speakers, even if the amount of contact was low. Together the combination of factors specific to the stimuli and participant groups used in these studies, relative to past experiments, may account for the unusually high accuracy rate in non-native speech recognition for both experienced and inexperienced listeners, and should be addressed in future work.

To investigate these questions, follow-up studies should more faithfully replicate the stimuli used in Baese-Berk et al.’s (2013) generalization study, recruiting speakers of the same accents and with comparable intelligibility levels to record the Bamford-Kowal-Bench (BKB) sentences which were used in their study. Intelligibility should be assessed in the same way as in NUFAESD, with participants judging sentences embedded in background noise at +5 dB SNR. Consistency with current experiments in recruitment from the same colleges in New York state, however, should address whether the increased accuracy is due to aspects of the stimuli or because the participant groups have more exposure in general, including those classified as inexperienced, than those in previous studies. In addition, it would be worthwhile to recruit participants who truly have no exposure to non-native speech, though this might be difficult from a college campus, to further investigate the questions asked in these studies.

Another drawback of the current methodology is the limited amount of demographic and social information that was solicited from participants for the sake of experiment duration. The Social Network Questionnaire did ask participants to report on their interactions with non-native speakers they knew personally or interacted with regularly, but it did not ask about other potential
experience with L2 accents, like media consumption or detailed travel experience (though listeners were asked on the Demographic Survey to list places they’d lived, including extended trips abroad). Extensive exposure to speech differing in accent from one’s own has been linked to improved comprehension of that accent, though only, to my knowledge, for native regional varieties (Adank et al. 2009; Evans & Iverson, 2004). It is therefore possible that by not soliciting this information from participants, differences in exposure to L2 accented speech were overlooked.

In addition to more detailed surveys of past experience with L2 speech, further research would benefit from incorporating more measures of potential social differences between groups. PPVT was selected as a means of testing for such differences because increased vocabulary size has been linked to greater comprehension of speech in noise (Banks et al., 2015; Bent et al., 2016; Tamati et al., 2013). It is possible there are other differences between participant groups in these studies that could impact their comprehension of L2-accented speech in noise. The control group in Experiment 2, along with 11 participants in the high-exposure condition, were recruited from a college in upstate New York. While no differences were found in vocabulary abilities between these participants and those recruited from New York City, either as freshmen or as high-exposure listeners, there are differences in the two schools’ average SAT scores for admitted students, with the average score for the NYC school being higher. No effect of improved comprehension of speech in noise has been found for IQ or education level beyond effects of vocabulary size, to my knowledge.

Literature on adaptation to difficult speech have found correlations between performance and other cognitive measures that could be illuminating in future work. (Banks et al., 2015) exposed participants to an artificial accent embedded in speech-shaped background noise.
Participants additionally completed a battery of cognitive ability tests, including Stroop, vocabulary knowledge, and working memory. The Stroop task requires participants to name the colour of the ink words were written in, which were incongruent with the word (e.g. blue written in red ink would require “red” as a response). Results indicate that listeners did adapt to the unfamiliar accents, and that Stroop scores significantly predicted better and faster adaptation. Vocabulary also predicted better performance, while working memory had a smaller effect. A benefit has been found across age groups for musical training (Parbery-Clark, Skoe, & Kraus, 2009; Parbery-Clark, Skoe, Lam, & Kraus, 2009; Strait, Parbery-Clark, Hittner, & Kraus, 2012; Talarico et al., 2007), and for executive function (Adank & Janse, 2010; Erb, Henry, Eisner, & Obleser, 2012; Huyck & Johnsrude, 2012; Janse & Adank, 2012). The purpose of including the 11 participants recruited from upstate New York in the high-exposure condition of Experiment 2 was precisely to ensure differences found between experimental groups in the three experiments and the control group (the low-exposure group in Experiment 2) were not due to differences due to recruitment location. Future work, however, should include a larger battery of assessments and questionnaires so as to be able to rule out other factors that could contribute to improved comprehension.

5.5 Conclusion

Previous research suggests listeners’ expectations of the speech signal can affect their comprehension and perception of unfamiliar voices. The literature has also demonstrated, however, that people can improve their comprehension of even difficult speech through high-variability exposure. Through a series of three experiments, this dissertation contributes several key findings to the growing body of literature investigating adaptation and generalization to
unfamiliar talkers. Consistent with past work (Baese-Berk et al., 2013), listeners who were trained on five unrelated non-native (L2) accented talkers had more accurate comprehension of both a novel talker of an accent they had been trained on, and a novel talker of an unfamiliar accent. These data provide further evidence that experience with a diversity of L2 accents facilitates generalization.

Experiment 2 tested whether this finding could be replicated with real-world exposure to L2-accented speech, rather than the artificial laboratory training that is common to attunement research. Individuals with greater lifetime experience listening to non-native accented speech, particularly from family, friends, and in their childhood community, were found to have more accurate comprehension of two unfamiliar L2 talkers compared to listeners with less prior exposure to non-native speech.

The third experiment investigated whether exposure to multiple different accents of unfamiliar English varieties would provide listeners with enough variability of experience to facilitate generalization to non-native voices. Training on five different native regional varieties of English did not facilitate comprehension of unfamiliar non-native talkers, however, and no difference was found compared to a control group trained on five speakers of American English. This suggests it is familiarity with the intra-talker variation characteristic of non-native speech that aids in generalization to other L2 talkers, and not just experience with inter-talker variation that differs from the listener’s own dialect.

Finally, people with more negative biases towards non-native speech were less accurate in their transcriptions of unfamiliar L2 talkers. This pattern was observed for all participant groups in all three experiments, with the exception of individuals who had extensive prior lifetime
experience interacting with non-native speakers. Though this group of listeners still exhibited the same range of biases as other conditions, biases were not found to be predictive of comprehension accuracy.
References


Appendix

Social Network Questionnaire

1. a. How many [friends, coworkers, house or roommates, neighbors, family members, teaching assistants, professors] do you have/have you had who are non-native speakers of English?

b. What is their native language? Guess or be general if you are not sure. Type "n/a" if you don't know any non-native speakers in this category.

2. When you were in high school, how many friends did you have who were non-native speakers of English?

3. When you were young (elementary school), how many friends did you have who were non-native speakers of English?

4. In your high school, were at least 30% of students non-native speakers of English?

5. In your elementary school, were at least 30% of students non-native speakers of English?

6. Do you think 30% or more of the people you regularly interacted with (either close to you or in customer service positions) when you were younger (elementary or high school age) were non-native speakers of English?
Statistical Output Tables

Experiment 1

Full model

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
Family: binomial (logit)
Formula: Accuracy ~ Condition + IAT + Language + (1 + Language | Participant) + (1 | Sentence)
Data: repncontrol

AIC BIC logLik deviance df.resid
2425.8 2481.2 -1204.9 2409.8 7540

Scaled residuals:
Min    1Q   Median    3Q   Max
-12.2991 0.0620 0.1198 0.2321 1.0887

Random effects:
Groups Name            Variance Std.Dev. Corr
Participant (Intercept) 0.5810   0.7622
Language Russian 0.5712   0.7558 -0.58
Sentence (Intercept) 2.5074   1.5835

Number of obs: 7548, groups: Participant, 74; Sentence, 30

Fixed effects:
     Estimate Std.Error z value Pr(>|z|)
(Intercept)     4.4634  0.4753   9.391  < 2e-16 ***
ConditionReplication 0.5709  0.1894   3.014  0.002579 **
IAT             0.9557  0.2600   3.676  0.000237 ***
LanguageRussian -0.6959  0.6285  -1.107    0.268207

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
     (Intr) ExprmR IAT
ConditionRplc -0.208
IAT            0.131
LanguageRssn -0.684  0.003

Base model vs. fixed effect of Condition

Models:
  base_expl: Accuracy ~ 1 + (1 | Participant)
  expl: Accuracy ~ Condition + (1 | Participant)

      Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
base_expl  2 2853.0 2866.8 -1424.5 2849.0
expl       3 2846.3 2867.1 -1420.2 2840.3 8.6212   1.000000

155
2 fixed effects: Condition + IAT

Models:

evaluated models:

\text{expl1}: \text{Accuracy} \sim \text{Condition} + (1 \mid \text{Participant})
\text{iatexpl1}: \text{Accuracy} \sim \text{Condition} + \text{IAT} + (1 \mid \text{Participant})

\begin{tabular}{llllll}
Df & AIC & BIC & logLik & deviance & Chisq & Chi Df & Pr(>Chisq) \\
\hline
expl1 & 3 & 2846.3 & 2867.1 & -1420.2 & 2840.3 \\
iatexpl1 & 4 & 2835.5 & 2863.2 & -1413.8 & 2827.5 & 12.841 & 1 & 0.0003391 *** \\
\end{tabular}

3 non-significant fixed effects: Condition, IAT, Predictability/Position/Recruitment Location

Models:

\text{iatexpl1}: \text{Accuracy} \sim \text{Condition} + \text{IAT} + (1 \mid \text{Participant})
\text{posiatexpl}: \text{Accuracy} \sim \text{Condition} + \text{IAT} + \text{Predictability} + (1 \mid \text{Participant})

\begin{tabular}{llllll}
Df & AIC & BIC & logLik & deviance & Chisq & Chi Df & Pr(>Chisq) \\
\hline
iatexpl1 & 4 & 2835.5 & 2863.2 & -1413.8 & 2827.5 \\
posiatexpl & 5 & 2837.4 & 2872.0 & -1413.7 & 2827.4 & 0.1201 & 1 & 0.7289 \\
\end{tabular}

Models:

\text{iatexpl1}: \text{Accuracy} \sim \text{Condition} + \text{IAT} + (1 \mid \text{Participant})
\text{posiatexpl}: \text{Accuracy} \sim \text{Condition} + \text{IAT} + \text{Position} + (1 \mid \text{Participant})

\begin{tabular}{llllll}
Df & AIC & BIC & logLik & deviance & Chisq & Chi Df & Pr(>Chisq) \\
\hline
iatexpl1 & 4 & 2835.5 & 2863.2 & -1413.8 & 2827.5 \\
posiatexpl & 6 & 2839.3 & 2880.9 & -1413.7 & 2827.3 & 0.166 & 2 & 0.9204 \\
\end{tabular}

Models:

\text{iatexpl1}: \text{Accuracy} \sim \text{Condition} + \text{IAT} + (1 \mid \text{Participant})
\text{coniexpl}: \text{Accuracy} \sim \text{Condition} + \text{IAT} + \text{Location} + (1 \mid \text{Participant})

\begin{tabular}{llllll}
Df & AIC & BIC & logLik & deviance & Chisq & Chi Df & Pr(>Chisq) \\
\hline
iatexpl1 & 4 & 2835.5 & 2863.2 & -1413.8 & 2827.5 \\
coniexpl & 5 & 2837.1 & 2871.8 & -1413.6 & 2827.1 & 0.3731 & 1 \\
\end{tabular}

\text{iatexpl1} \quad \text{coniexpl} \quad 0.5413

3 fixed effects: Condition, IAT, Test Language

Models:

\text{iatexpl1}: \text{Accuracy} \sim \text{Condition} + \text{IAT} + (1 \mid \text{Participant})
\text{lgiatexpl}: \text{Accuracy} \sim \text{Condition} + \text{IAT} + \text{Language} + (1 \mid \text{Participant})

\begin{tabular}{llllll}
Df & AIC & BIC & logLik & deviance & Chisq & Chi Df & Pr(>Chisq) \\
\hline
iatexpl1 & 4 & 2835.5 & 2863.2 & -1413.8 & 2827.5 \\
lgiatexpl & 5 & 2812.0 & 2846.7 & -1401.0 & 2802.0 & 25.498 & 1 & 4.428e-07 *** \\
\end{tabular}

Adding random effect of sentence

Models:

\text{lgiatexpl}: \text{Accuracy} \sim \text{Condition} + \text{IAT} + \text{Language} + (1 \mid \text{Participant})
\text{randomeffects}: \text{Accuracy} \sim \text{Condition} + \text{IAT} + \text{Language} + (1 \mid \text{Participant}) + \text{randomeffects} (1 \mid \text{Sentence})
Random slopes for language by participant

Models:
randomeffects: Accuracy ~ Condition + IAT + Language + (1 | Participant) +
randomeffects:   (1 | Sentence)
randomslopes: Accuracy ~ Condition + IAT + Language + (1 + Language | Participant) + (1 | Sentence)

Experiment 2

Full model: reference level = high exposure

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
  Family: binomial  (logit)
  Formula: Accuracy ~ Condition + IAT + Language + (1 | Participant) + (1 | Sentence)
  Data: lifetime_new

AIC      BIC   logLik deviance df.resid
1998.3   2046.3 -992.2   1984.3     6929

Scaled residuals:
  Min       1Q   Median       3Q      Max
-13.5681   0.0556   0.1085   0.2134   1.3253

Random effects:
  Groups      Name        Variance Std.Dev.
  Participant (Intercept) 0.172    0.41
  Sentence    (Intercept) 2.729    1.6519
Number of obs: 6936, groups: Participant, 68; Sentence, 30

Fixed effects:   Estimate Std. Error z value Pr(>|z|)
(Intercept)      5.6396     0.5022 11.229  < 2e-16 ***
Conditionlowexposure -1.2666     0.1906 -6.647 2.99e-11 ***
ConditionSUNY_notraining -1.6808     0.2838 -5.922 3.19e-09 ***
IAT                  0.8958     0.2445  3.664 0.000249 ***
LanguageRussian   -0.6857     0.6497 -1.055 0.291278

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
  (Intr) Cndtnl CSUNY_ IAT
CndtnlIwxpsr -0.255
CndtnSUNY_n -0.185 0.439  
IAT 0.090 0.101 -0.074  
LanguagRssn -0.677 0.003 0.003 -0.002

Reference level = low exposure

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
Family: binomial (logit)
Formula: Accuracy ~ Condition + IAT + Language + (1 | Participant) + (1 | Sentence)
Data: lifetime_new

AIC      BIC   logLik deviance df.resid
1998.3   2046.3 -992.2   1984.3     6929

Scaled residuals:
Min       1Q   Median       3Q      Max
-13.5681  0.0556  0.1085  0.2134 1.3253

Random effects:
Groups     Name        Variance Std.Dev.
Participant (Intercept) 0.1721   0.4148
Sentence    (Intercept) 2.7292   1.6520
Number of obs: 6936, groups: Participant, 68; Sentence, 30

Fixed effects: 
(Intercept)                4.3731     0.4898   8.929  < 2e-16 ***
Conditionhighexposure      1.2666     0.1906   6.647 2.99e-11 ***
ConditionSUNY_notraining  -0.4141     0.2635  -1.572 0.115987
IAT                       0.8958     0.2445   3.664 0.000249 ***
LanguageRussian           -0.6856     0.6499  -1.055 0.291407

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
   (Intr) Cndtnh CSUNY_ IAT
Cndtnghxps -0.128
CndtnSUNY_n -0.113 0.251
IAT           0.131 -0.101 -0.153
LanguagRssn  -0.693 -0.003 0.001 -0.002

Base model vs. fixed effect of Condition

Models:
base_exp2: Accuracy ~ 1 + (1 | Participant)
exp2: Accuracy ~ Condition + (1 | Participant)

<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>base_exp2</td>
<td>2</td>
<td>2377.1</td>
<td>2390.8</td>
<td>-1186.6</td>
<td>2373.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp2</td>
<td>4</td>
<td>2340.2</td>
<td>2367.6</td>
<td>-1166.1</td>
<td>2332.2</td>
<td>40.867</td>
<td>2</td>
<td>1.336e-09   ***</td>
</tr>
</tbody>
</table>
2 fixed effects: Condition + IAT

Models:
exp2: Accuracy ~ Condition + (1 | Participant)
iatexp2: Accuracy ~ Condition + IAT + (1 | Participant)

<table>
<thead>
<tr>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>exp2</td>
<td>4</td>
<td>2340.2</td>
<td>2367.6</td>
<td>-1166.1</td>
<td>2332.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>iatexp2</td>
<td>5</td>
<td>2330.3</td>
<td>2364.5</td>
<td>-1160.2</td>
<td>2320.3</td>
<td>11.937</td>
<td>0.0005503 ***</td>
</tr>
</tbody>
</table>

3 non-significant fixed effects: Condition, IAT, Predictability/Position/Recruitment Location

Models:
iatexp2: Accuracy ~ Condition + IAT + (1 | Participant)
lociatexp2: Accuracy ~ Condition + IAT + Location + (1 | Participant)
posiatexp2: Accuracy ~ Condition + IAT + Position + (1 | Participant)
prediatspeech: Accuracy ~ Condition + IAT + Predictability + (1 | Participant)

<table>
<thead>
<tr>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>iatexp2</td>
<td>5</td>
<td>2330.3</td>
<td>2364.5</td>
<td>-1160.2</td>
<td>2320.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lociatexp2</td>
<td>6</td>
<td>2332.2</td>
<td>2373.2</td>
<td>-1160.1</td>
<td>2320.2</td>
<td>0.1637</td>
<td>0.6858</td>
</tr>
<tr>
<td>posiatexp2</td>
<td>7</td>
<td>2334.2</td>
<td>2381.2</td>
<td>-1160.1</td>
<td>2320.2</td>
<td>0.0953</td>
<td>0.9535</td>
</tr>
<tr>
<td>prediatspeech</td>
<td>8</td>
<td>2339.8</td>
<td>2380.8</td>
<td>-1158.9</td>
<td>2317.8</td>
<td>0.0953</td>
<td>0.9535</td>
</tr>
</tbody>
</table>

3 fixed effects: Condition, IAT, Test Language

Models:
iatexp2: Accuracy ~ Condition + IAT + (1 | Participant)
lgiatexp2: Accuracy ~ Condition + IAT + Language + (1 | Participant)

<table>
<thead>
<tr>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>iatexp2</td>
<td>5</td>
<td>2330.3</td>
<td>2364.5</td>
<td>-1160.2</td>
<td>2320.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lgiatexp2</td>
<td>6</td>
<td>2313.6</td>
<td>2354.6</td>
<td>-1150.8</td>
<td>2301.6</td>
<td>18.754</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
</tbody>
</table>

Adding random effect of sentence

Models:
lgiatexp2: Accuracy ~ Condition + IAT + Language + (1 | Participant) + (1 | exp2sent: Sentence)

<table>
<thead>
<tr>
<th>Df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
<th>Chisq</th>
<th>Chi Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>lgiatexp2</td>
<td>6</td>
<td>2313.6</td>
<td>2354.6</td>
<td>-1150.8</td>
<td>2301.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp2sent</td>
<td>7</td>
<td>1998.3</td>
<td>2046.2</td>
<td>-992.17</td>
<td>1984.3</td>
<td>317.22</td>
<td>1 &lt; 2.2e-16 ***</td>
</tr>
</tbody>
</table>

Random slopes for language by participant

Models:
Experiment 3

Full model: reference level = Exp2 Control (American English trained)

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
Family: binomial  (logit)
Formula: Accuracy ~ Experiment + IAT + Language + (1 + Language | Participant) + (1 | Sentence)
Data: nvdata

AIC      BIC   logLik deviance df.resid
3922.4   3988.7 -1952.2   3904.4    11721
Scaled residuals:
    Min       1Q   Median       3Q      Max
-19.7532   0.0638   0.1262   0.2501  1.0454

Random effects:
  Groups      Name            Variance Std.Dev. Corr
  Participant (Intercept)     0.4344   0.6591
    LanguageRussian 0.3187   0.5645 -0.58
  Sentence (Intercept)     2.5743   1.6045
Number of obs: 11730, groups:  Participant, 115; Sentence, 30

Fixed effects:
        Estimate Std. Error z value Pr(>|z|)
(Intercept) 4.4385     0.4596   9.657  < 2e-16 ***
ExperimentNative 0.1174     0.1710   0.687  0.49229
ExperimentReplication 0.4800     0.1620   2.964  0.00304 **
IAT           0.8406     0.1929   4.358  1.31e-05 ***
LanguageRussian -0.6187     0.6186  -1.000   0.31722

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
    (Intr) ExprmN ExprmRplc IAT
ExpermntNtv -0.157
ExprmrntRplc 0.169  0.430
IAT          0.125 -0.095 -0.083
LanguageRssn -0.690 -0.010  0.004  0.014

Full model: reference level = Native varieties trained
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
Family: binomial (logit)
Formula: Accuracy ~ Experiment + IAT + Language + (1 + Language | Participant) + (1 | Sentence)
Data: nvdata

AIC      BIC   logLik deviance df.resid
3922.4   3988.7 -1952.2   3904.4    11721

Scaled residuals:
    Min       1Q   Median       3Q      Max
-19.7531   0.0638   0.1262   0.2501   1.0455

Random effects:
  Groups      Name    Variance  Std.Dev.  Corr
  Participant (Intercept)  0.4344   0.6591
  LanguageRussian  0.3187   0.5645  -0.58
  Sentence    (Intercept)  2.5742   1.6044
Number of obs: 11730, groups:  Participant, 115; Sentence, 30

Fixed effects: Estimate Std. Error z value Pr(>|z|)
(Intercept)             4.5559     0.4647   9.805  < 2e-16 ***
ExperimentLifetime        -0.1174     0.1710  -0.686   0.4924
ExperimentReplication     0.3626     0.1780   2.038   0.0416 *
IAT                      0.8406     0.1929   4.359 1.31e-05 ***
LanguageRussian        -0.6187     0.6186  -1.000   0.3173
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
 (Intr) ExprmntL ExprmntRplc IAT
ExprmntLftm   -0.213
ExprmntRplc   -0.213  0.570
IAT           0.088  0.095  0.015
LanguageRssn -0.686  0.011  0.014  0.014

Base model vs. fixed effect of Condition

Models:
 base_exp3: Accuracy ~ 1 + (1 | Participant)
 exp3: Accuracy ~ Experiment + (1 | Participant)

Df  AIC  BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
base_exp3 2 4608.1 4622.8 -2304.0  4604.1
exp3      4 4601.1 4630.5 -2296.5  4593.1 11.008  2   0.004071 **

2 fixed effects: Condition + IAT

Models:
 exp3: Accuracy ~ Experiment + (1 | Participant)
 iatexp3: Accuracy ~ Experiment + IAT + (1 | Participant)
3 non-significant fixed effects: Condition, IAT, Predictability/Position/Recruitment/Location

Models:
iatexp3: Accuracy ~ Experiment + IAT + (1 | Participant)
lociatexp3: Accuracy ~ Experiment + IAT + Condition + (1 | Participant)

Models:
iatexp3: Accuracy ~ Experiment + IAT + (1 | Participant)
posiatexp3: Accuracy ~ Experiment + IAT + Position + (1 | Participant)

Models:
iatexp3: Accuracy ~ Experiment + IAT + (1 | Participant)
prediatexp3: Accuracy ~ Experiment + IAT + Predictability + (1 | Participant)

Models:
iatexp3: Accuracy ~ Experiment + IAT + (1 | Participant)
lgiatexp3: Accuracy ~ Experiment + IAT + Language + (1 | Participant)
Adding random effect of sentence

Models:
lgiatexp3: Accuracy ~ Experiment + IAT + Language + (1 | Participant)
exp3sent: Accuracy ~ Experiment + IAT + Language + (1 | Participant) +
         exp3sent:     (1 | Sentence)

Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
lgiatexp3  6 4562.7 4606.9 -2275.3   4550.7
exp3sent   7 3924.5 3976.1 -1955.2   3910.5 640.22      1 < 2.2e-16 ***

Random slopes for language by participant

Models:
exp3sent: Accuracy ~ Experiment + IAT + Language + (1 | Participant) +
         exp3sent:     (1 | Sentence)
exp3slopes: Accuracy ~ Experiment + IAT + Language + (1 + Language | Participant) +
           exp3slopes:     (1 | Sentence)

Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
exp3sent    7 3924.5 3976.1 -1955.2   3910.5
exp3slopes  9 3922.4 3988.7 -1952.2   3904.4 6.051 2    0.04853 *

Peabody Picture Vocabulary Test

Summary stats per participant group

<table>
<thead>
<tr>
<th>Condition</th>
<th>Experiment</th>
<th>Training voices</th>
<th>PPVT (μ)</th>
<th>PPVT (σ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>207.66</td>
<td>6.55</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>206.58</td>
<td>8.99</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>No training</td>
<td>206.63</td>
<td>9.49</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>206.45</td>
<td>7.17</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>206.62</td>
<td>9.95</td>
</tr>
</tbody>
</table>

Score per participant

<table>
<thead>
<tr>
<th>Condition</th>
<th>Experiment</th>
<th>Training voices</th>
<th>PPVT</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>212</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>204</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>190</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>209</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>213</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>208</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>195</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>205</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>213</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>212</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>207</td>
</tr>
<tr>
<td>Location</td>
<td>Type</td>
<td>Speakers</td>
<td>Number</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>----------</td>
<td>--------</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>202</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>210</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>203</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>211</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>200</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>215</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>210</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>217</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>210</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>212</td>
</tr>
<tr>
<td>NYC</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>210</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>206</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>216</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>213</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>200</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>208</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>213</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>217</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>213</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>215</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>201</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>201</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>200</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>217</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>200</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>210</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>210</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>200</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>203</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>215</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>215</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>191</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>187</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>198</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>199</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>189</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>208</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE</td>
<td>213</td>
</tr>
<tr>
<td>Location</td>
<td>Lifetime</td>
<td>Speaker Type</td>
<td>Number</td>
</tr>
<tr>
<td>------------------------</td>
<td>----------</td>
<td>--------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Upstate</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>220</td>
</tr>
<tr>
<td>Upstate (high exposure)</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>197</td>
</tr>
<tr>
<td>Upstate (high exposure)</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>215</td>
</tr>
<tr>
<td>Upstate (high exposure)</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>212</td>
</tr>
<tr>
<td>Upstate (high exposure)</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>209</td>
</tr>
<tr>
<td>Upstate (high exposure)</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>196</td>
</tr>
<tr>
<td>Upstate (high exposure)</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>209</td>
</tr>
<tr>
<td>Upstate (high exposure)</td>
<td>Lifetime</td>
<td>5 MAE speakers</td>
<td>200</td>
</tr>
<tr>
<td>Upstate (no training)</td>
<td>Lifetime</td>
<td>No training</td>
<td>216</td>
</tr>
<tr>
<td>Upstate (no training)</td>
<td>Lifetime</td>
<td>No training</td>
<td>210</td>
</tr>
<tr>
<td>Upstate (no training)</td>
<td>Lifetime</td>
<td>No training</td>
<td>198</td>
</tr>
<tr>
<td>Upstate (no training)</td>
<td>Lifetime</td>
<td>No training</td>
<td>207</td>
</tr>
<tr>
<td>Upstate (no training)</td>
<td>Lifetime</td>
<td>No training</td>
<td>212</td>
</tr>
<tr>
<td>Upstate (no training)</td>
<td>Lifetime</td>
<td>No training</td>
<td>218</td>
</tr>
<tr>
<td>Upstate (no training)</td>
<td>Lifetime</td>
<td>No training</td>
<td>202</td>
</tr>
<tr>
<td>Upstate (no training)</td>
<td>Lifetime</td>
<td>No training</td>
<td>190</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>209</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>208</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>210</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>202</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>203</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>199</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>195</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>203</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>196</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>194</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>218</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>194</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>202</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>202</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>211</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>199</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>214</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>213</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>216</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>211</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>207</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>202</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>205</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>212</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>210</td>
</tr>
<tr>
<td>Location</td>
<td>Language Type</td>
<td>Speaker Count</td>
<td>Code</td>
</tr>
<tr>
<td>----------</td>
<td>---------------</td>
<td>---------------</td>
<td>--------</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>212</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>213</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>206</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>215</td>
</tr>
<tr>
<td>NYC</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>219</td>
</tr>
<tr>
<td>Upstate</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>200</td>
</tr>
<tr>
<td>Upstate</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>188</td>
</tr>
<tr>
<td>Upstate</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>195</td>
</tr>
<tr>
<td>Upstate</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>208</td>
</tr>
<tr>
<td>Upstate</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>213</td>
</tr>
<tr>
<td>Upstate</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>196</td>
</tr>
<tr>
<td>Upstate</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>199</td>
</tr>
<tr>
<td>Upstate</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>207</td>
</tr>
<tr>
<td>Upstate</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>222</td>
</tr>
<tr>
<td>Upstate</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>212</td>
</tr>
<tr>
<td>Upstate</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>210</td>
</tr>
<tr>
<td>Upstate</td>
<td>Replication</td>
<td>5 L2 speakers</td>
<td>221</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>214</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>216</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>199</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>172</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>200</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>203</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>196</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>206</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>212</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>204</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>198</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>202</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>218</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>212</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>208</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>201</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>220</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>210</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>214</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>215</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>219</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>210</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>203</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>212</td>
</tr>
<tr>
<td>-----</td>
<td>------------------</td>
<td>---------------------</td>
<td>-----</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>197</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>195</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>214</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>216</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>199</td>
</tr>
<tr>
<td>NYC</td>
<td>Native Varieties</td>
<td>5 regional speakers</td>
<td>210</td>
</tr>
</tbody>
</table>