The Effects of Explicit Knowledge of and Implicit Attitudes about Race on Adult Perceptions of Children's Speech

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Andrea Lynn Christy

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS

Benjamin R. Munson, Adviser

August 2010

Acknowledgements

While this thesis bears my name as the sole author, no project of this magnitude comes to fruition without the efforts of an entire community. I owe so much to the group of people who so graciously assisted me in bringing this project into being.

First of all I would like to thank my adviser, Benjamin Munson, for his unwavering support and incredible work ethic. He has been a model of generosity in all ways: intellectually, professionally, and emotionally. I would not have been able to get to this point as a student researcher and a future Speech-Language Pathologist without his guidance and advice. Thank you to my wonderfully understanding and accommodating committee members, Mark DeRuiter and Nancy Stenson. Your enthusiasm for this project and your insightful suggestions have been inspirational and invaluable.

Thank you to Molly Babel and Jan Edwards for rich and engaging email conversations and for your generosity with your research. We would not have the data from the IAT without Molly Babel's IAT protocol.

Thank you to the faculty in the Speech-Language-Hearing Sciences department who have assisted me in the completion of this project either indirectly or directly through lectures, articles and chance conversations that peaked my interest and guided the development of this project. Your knowledge and enthusiasm for research and the discipline have been invaluable. Special thanks to Leslie Glaze from whom I learned so much about research and statistics as a teaching assistant for Rate Your World.

Thank you to the Center for Early Education and Development at the University of Minnesota for the use of royalty free photos of children used in the project.

I could not have accomplished this task without the incredible support of the "Labbies" in Shevlin 5. Thank you to: Lauren Derksen, Eden Kaiser, Marie Meyer, Kari Urberg-Carlson, Hannah Julien, Julie Johnson, Laura Crocker, Renata Solum, Sarah Schellinger, Maura Arnoldy, Sarah Mans, Anna Schnurrer, and Muhammad Abdurrahman. Our weekly meetings provided much needed support and feedback. To Lauren, Marie, Eden, Renata, and Kari: I could not have accomplished the data collection without you. I am in your debt.

Thank you to everyone who participated in this study. Without you, this project would not exist.

This research project was generously funded in part by National Science Foundation grant number BCS 0729277.

Finally, I am so grateful for the love and support I have received from my husband, Steve, and my family and friends. I am truly lucky to have found such wonderful people with whom I can share my passions and my life. Steve, your patience, kindness, and steadfastness made this project possible. There is no better editor out there! Thank you Mom, Sherri, Kerry Medenwald, Heather Wood-Davila, Barbara Leonard, and Nicole Grunzke for keeping me on an even-keel through this process. People like you help make this world a better place for all of us.

Thank you all for believing in me and this project!

Dedication

To my husband, my cats, my mom…

and to all people who are striving to attain or regain their highest level of communication, who deserve the best possible services no matter their cultural background or social status.

Table of Contents

List of Tables

List of Figures

1. Introduction

Speech is a highly variable phenomenon. It is nearly axiomatic that variation in the acoustic forms of words is caused by many factors. In perception, listeners must relate a highly variable signal to an invariant representation of sounds and words in long-term memory. For most, if not all speech sounds, there is no invariant acoustic property that listeners can attend to when perceiving speech; this is known as the *invariance* problem (Perkell & Klatt, 1986). Consequently, listeners must apply their knowledge of the sources of acoustic variation when perceiving the acoustically variable speech signal. Researchers have yet to observe a consistent correlation between the acoustic signal produced by talkers and the linguistic elements of the signal that listeners perceive and interpret. Variability exists both between speakers and within an individual's speech production. The anatomy of the larynx and of the vocal tract constrains the acoustic characteristics of speech: individuals with smaller larynxes and shorter vocal tracts produce speech that has higher perceived pitch and higher resonant frequencies than individuals with larger larynxes and longer vocal tracts. Hence, listeners must learn to interpret the acoustic signal that a talker produces relative to the overall range of frequencies that they produce. In running speech, coarticulatory processes influence sounds' acoustic detail, creating a range of acoustic signals that are perceived as the same phoneme depending on the phonemic context (Lindblom, 1990). Similarly, variations in rate, stress, and accent contribute to the high complexity of speech, the perception of which is context-dependent. Variation between talkers is similarly daunting. Differences in pronunciation across ages, genders, and dialect groups leads to even more variation in

speech sounds. Yet listeners perceive and interpret speech in a surprisingly efficient and effortless manner. Arguably, listeners who apply their knowledge of the sources of variance (i.e., the social, linguistic, and anatomical factors that condition variation in speech sounds) can perform these the speech-decoding processes better than those who cannot apply this knowledge, or who do not have a full set of knowledge.

Numerous studies have been conducted to help better understand how listeners resolve the dilemma posed by the invariance problem. Visual information has been shown to dramatically affect the perception of speech at the phonemic and lexical level. The *McGurk effect* (McGurk & McDonald, 1976) demonstrated the direct impact visual information has on listeners' perception of auditory stimuli. Listeners presented with a video representation of a person saying the syllable /ga/ or /ba/ while hearing the opposite syllable being produced actually perceived the syllable /da/, the phoneme /d/ being intermediary in articulatory placement between /g/ and /b/. Interestingly, this phenomenon occurs even when listeners are aware of the effect and of the actual identity of the stimulus (Walker, Bruce, & O'Malley, 1995). Miller's (1981) summary of research on the relationship between speaking rate and listener perception suggests that even though changes in speaking rate affect different components of speech in diverse ways, listeners are able compensate for these variations to a certain extent. Studies summarized by Miller showed that, when presented with speech segments, listeners judged tempo and phonetic structure at the same time and as interacting with each other. Prosody, therefore, constitutes a component of the context in which speech is perceived. Additionally, listeners regularly use lexical cueing to interpret grammatically ambiguous

instructions. For example, “put the apple on the towel...” may be referring to where one will find the apple or to where one is expected to put the apple. In this case, people used the visual referent to disambiguate the instructions after momentary confusion (Tanenhaus, 1995). Listeners also effortlessly employ emotive context to choose between ambiguous homophones. When asked to describe homophones, one of which carries a lexical affect paired with one that does not, listeners transcribed the homophone with affect more often when the audio stimulus was delivered with a congruous affect in the tone of voice. For example, the sequence [dɑɪ] was transcribed more often as *die* than *dye* when paired with a sad tone of voice. This suggests that listeners register emotional tone of voice and use it when assigning word meaning to homophones (Nygaard & Lunders, 2002).

Impact of Indexical Information on Speech Perception

This project is part of a larger body of inquiry concerned with how known or assumed indexical information (e.g. perceived gender, race, age, class, and nationality of origin) affects listener language perception and comprehension. A small but growing body of literature has examined the judgments made by listeners about speakers and their perceived social identity from speech production. For example, listeners judged a speaker’s intelligence differently, among many other factors, based upon talkers’ pronunciations of ING and the formality of the conversational setting (Campbell-Kibler, 2006). The production of “runnin’” instead of “running” in a speech sample elicited varied assumptions about the talker from different listeners including that the person was less intelligent and that the person was from the Southern United States. Recently, there

has been a surge of interest in investigating how listener-perceived or listener-assumed indexical information about the speaker affects speech perception and comprehension. This research has changed and continues to shape new views about the phenomena of speech perception. The next sections review these studies.

Regional Dialect

An influential study by Niedzielski (1999) examined the effect stereotypes about the speech of people with national orgins on speech perception. In that study, middle-class Detroit residents listened to a speech sample of a middle-class Detroit woman with a Northern Cities Chain Shift (NCCS) dialect, which is typical of middle-class Detroit residents, and includes raised-diphthong vowels (i.e., productions of the word *night* with the vowel [ʌɪ]). One group of listeners was told they were listening to a Canadian woman while the other group was informed the woman was from Detroit. Those who thought the speaker was Canadian matched speech tokens with raised-diphthongs to the Detroit targets (i.e., they identified [ʌɪ] tokens as from the word *night*), whereas those who thought she was from Detroit matched speech tokens with vowels closer to Standard American English dialect to the speech sample (i.e., they identified [ɑɪ] as *night*). Even though middle-class Detroit residents typically speak with a NCCS dialect, they perceived a dialect closer to Standard American English when told a speaker they were listening to was from Detroit. They did perceive the raised-diphthong when they believed they were listening to a Canadian speaker. The perception of dialect changed depending on where the listener thought the woman was from. Two findings from this study have particularly important implications. First, perception is significantly affected

by expectations raised by stereotypical beliefs about an aspect of a speaker's identity. Second, social stigmatization of non-standard dialects may influence Detroit listeners to perceive a fellow Detroit resident's dialect as Standard American English.

In a more recent study of the influence of listener stereotypes about national identity on speech perception, Hay and colleagues (Hay, 2006; Drager & Hay, 2006) found a similar relationship between perceived identity and phoneme identification. Here, the variable of interest was the high-front vowel /ɪ/, which is pronounced very differently in New Zealand and Australia. This experiment used a priming methodology to lead listeners to believe that the speaker was from New Zealand or Australia. New Zealander listeners changed their perception of that vowel based solely on whether they identified the speaker to be a New Zealander or an Australian as gauged by self-report. Surprisingly, this perception shift occurred even though almost all of those given the "Australian" priming identified the speaker as being from New Zealand. Even more striking, in a subsequent study the mere presence of a stuffed kangaroo toy or a stuffed kiwi bird toy in the testing room resulted in a similar perception shift, suggesting that speech perception can depend on contextual cues to the national origin of the talker.

Age and Class

Drager (2005) performed an experiment that attempted to ascertain the effect of perceived age on speech perception. Drager's study focused on a set of vowels whose pronunciation differs between older and younger speakers as the result of a sound change in progress occurring across New Zealand. In younger speakers, the pronunciation of vowels in words like TRAP is becoming more raised (i.e., sounding similar to the vowel

in DRESS). Older speakers pronounce the DRESS and TRAP vowels similarly to American English speakers. Drager also used a priming methodology, combining audio tokens produced by older and younger talkers with pictures of older or younger talkers. Listeners were also asked to judge the age of the speakers. While the effect of picture-age and voice-age on phoneme identification did not achieve statistical significance using conventional criteria, an important pattern in listeners' judgments of the vowels was noted. Regardless of which speaker the listeners judged as older, a greater percentage of the tokens spoken by the "younger" speaker were judged as being the raised variant, even though the vowels were acoustically identical. Thus the perceived age of the speaker caused the listeners to judge word tokens with identical vowels as different.

Hay, Warren and Drager (2006) capitalized on a similar vowel shift occurring in New Zealand. New Zealand English is non-rhotic, the /r/ being replaced by a diphthong with a schwa offglide. The diphthong in words like SQUARE is merging in younger speakers and working class speakers into a pronunciation closer to the diphthong in words like NEAR. This diphthong merge causes certain minimal word pairs to become ambiguous without further contextual clues when spoken by younger New Zealanders, for example *pair* and *peer* sound the same. This experiment used a list of such minimal word pairs coupled with photos that participants were asked to imagine represented the speakers. Researchers presented one group with photos that differentiated the speakers by age. The other group saw photos that connoted a speaker as middle class or working class.

This work also showed an interesting association between listeners' own production and how they perceived others' speech. Listeners who themselves still speak with two distinct diphthongs (*pair* and *peer* sound different) were significantly affected by presumed age of the speaker. When listening to samples paired with photos of an older speaker, their accuracy in a forced choice task between two minimal word pairs was high. However, when listeners presumed the speaker to be younger, error rates increased significantly as the listeners appeared to treat the productions as ambiguous as they would expect from a younger speaker. Perceived social class also affected listeners' perception. Listeners whose own diphthongs were more merged increased their accuracy in the forced choice task as the social class of the photo increased. Interestingly, participants who produced distinct diphthongs showed the opposite trend. As the portrayed class of the people in the photos increased, so did the error rates. The researchers believed this discrepancy was caused by a lack of exposure to lower class speech. Possibly, the only exposure these listeners had to the vowel shift was in the speech of younger speakers, and thus associated this vowel shift with the age of the speaker and not with the class of the speaker.

Gender

In 1996, Strand and Johnson demonstrated a significant perceptive shift in fricatives (/s/ and /ʃ/) depending on the gender of a photo shown to listeners that purportedly represented the speaker. When listening to an ambiguous phoneme midway between /s/ and /ʃ/ acoustically listeners consistently reported perceiving /ʃ/ when they were shown a photograph of a person that had been deemed more stereotypically female

in a previous study. This follows known acoustic differences between male and female sibilant production, females producing a higher acoustic boundary between /s/ and /ʃ/ overall. An analogous result was found by Johnson, Strand, and D'Imperio (1999), who showed that perceived gender affected vowel labeling.

Race/Ethnicity

Rubin (1992) devised a study that examined the impact of foreign accented English spoken by Teaching Assistants on the comprehension of North American undergraduate students listening to a lecture. In one phase of the study a recording of lectures delivered by the same native speaker of Standard American English was paired with a photo of an Asian face or a Caucasian face projected onto a screen in the front of the room. Students who attended the lecture paired with an Asian face showed significantly lower comprehension scores and reported significantly higher judgments of the lecturer's pronunciation to be more foreign sounding or nonstandard than did those who listened to the lecture paired with the Caucasian face. In a more recent study, listeners were shown to use implicit knowledge of racial dialects to disambiguate sentences (Staum Cassasanto, 2008). One of the most salient and consistent markers of African American English dialect is the deletion of final stop consonants in final consonant clusters, causing words like *wind* to sound like standard American English pronunciations of *win*. In Staum Casasanto's study, participants listened to the recording of a phrase that contained a word that could be perceived as either having a reduced final cluster or not. For example, the syllable [mæs] was used in the phrase "The [mæs] probably lasted…" which could be interpreted as "mast" or "mass" depending on

contextual clarification. The stimulus was paired with the photo of an African American or a Caucasian. After hearing the phrase, listeners were asked to read a phrase that completed the sentence and judge whether or not the completed sentence made sense. One of the written stimuli made sense if the final consonant cluster was reduced: "The [mæs] probably lasted…through the storm." The other made sense if there was no cluster reduction: "The [mæs] probably lasted…an hour on Sunday." Response times were measured. Response times for the phrase that made sense with a presumed cluster reduction were faster when paired with an African American face (consistent with the assumption that the final consonant cluster was reduced from "mast" to "mas'" in African American English [AAE] dialect). Response times for the phrase that made sense with no presumed cluster reduction were also significantly faster when paired with a Caucasian face. In this case, listener implicit knowledge of racial dialects aided in faster linguistic processing. Critically for this investigation, Staum Casasanto showed that simply pairing an African American face with a speech token led listeners to believe that they were listening to a speaker of African American English. This is interesting because simply being African American does not guarantee that one will speak AAE. Hence, the listeners in Staum Casasanto's study appeared to over-generalize the relationship between people's race and their dialect.

Implications for the clinical practice of Speech-Language Pathology

The evidence for the impact of indexical or social information on speech perception continues to grow. Even when people consciously recognize social information, conflicting information from the environment can override the conscious

understanding and affect speech perception and comprehension as in the case discussed above of the New Zealanders who knew they were listening to fellow New Zealanders yet reported a marked perceptual shift in vowels towards an Australian dialect in the presence of an answer sheet with the word "Australian" across the top (Hay, et al., 2006). The implications of this kind of perceptual shift even in the presence of conflicting conscious knowledge brings to mind concerns about the impact on the practice of Speech Language Pathology wherein clinicians must use their own perception to evaluate a production as typical for a given dialect or as disordered. Do shifts in perception like those observed by Hay, Niedzielski, Staum Casasanto, and others affect how listeners might rate the accuracy of a given speech production by a child? Can explicit, conscious knowledge of dialectal differences override automatic perceptual shifts that occur in the presence of implicit biases? Do implicit biases cause perceptual shifts that confound the distinction between dialectical difference and disordered speech?

There exists little basic science to guide us in the effects implicit biases have on clinical practice especially in the assessment of children's speech. From the small number of studies we do have, evidence shows that biasing does impact adult perceptions of children's speech. Munson and Seppanen (2009) presented adult listeners with a set of narratives paired with photos of boys and girls. When asked to judge the quality of the narratives, listeners rated the exact same narratives as being of higher quality when paired with boys' photos than when paired with photos of girls. Munson and Seppanen posited that listeners expect that girls in general have more developed language skills than boys.

Thus, when listening to a narrative spoken presumably by a girl, their criteria were higher than when they thought they were listening to a boy.

Presumed age of a child has also been shown to influence adult perceptions of child's speech and ratings of accuracy of production. Two recent studies found small effects of perception and ratings of misarticulation of /s/ as /θ/. If listeners believed the child to be older, they were more likely to rate productions of /θ/ as accurate than if they believed the child to be younger (Munson, Edwards, Schellinger, Beckman, & Meyer, 2010; Schellinger, Edwards, & Munson, 2010).

The purpose of this experiment was to build our understanding of adult listeners' perceptions of children's speech, and how their perceptions might be affected by their perception of social attributes about the child, specifically, the race of the child. We examined the effects of presumed race of a speaker on the perception of a listener because the distinction between dialectal difference and disordered speech is particularly essential in clinical practice. In this case, we focused on African American English and Standard American dialects. This was accomplished by constructing an audiovisual speech perception experiment, in which audio speech tokens produced by children were paired with pictures of African American or Caucasian children. Specifically, we examined whether listeners had different criteria for speech-sound accuracy when sounds were paired with pictures of African American or Caucasian children's faces.

This study differs from previous research on this topic in a number of ways. First, this study systematically manipulated the 'accuracy' of children's speech by using speech tokens that were acoustically manipulated to be either clearly accurate or clearly

inaccurate. Two sets of audio speech tokens were used. The first set of tokens were /s/-initial words, acoustically manipulated to have either a clearly accurate /s/, or an /s/ that was clearly frontally misarticulated (i.e., had a /θ/ pronunciation). Frontally misarticulated /s/ is a common speech error in children, as shown in large-scale normative studies, such as that by Smit, Hand, Freilinger, Bernthal, and Bird (1990). The second set of tokens was comprised of words that end in /t/. These were acoustically manipulated to give the illusion of a glottalized or deleted final consonant. Again, deleted final consonants are common speech errors in children, as shown by studies like Smit et al. (1990). Moreover, the glottalization of word-final /t/ is also a characteristic of African American English, as shown in Thomas's (2007) review of the phonological characteristics of African American English. By using these two sets of stimuli, we were able to examine whether the influence of picture race on judgments of the accuracy of children's speech was equivalent for one variable that is known to differ between African American English speakers and Standard American English speakers (final consonant production) and one that is not (initial /s/ production).

Second, this study is innovative because it systematically measured listeners' experience perceiving children's speech. This was accomplished two ways. First, we had all individuals report the amount of time they spent listening to children. Second, we included two groups that we believed *a priori* would have different amounts of time perceiving children's speech: undergraduate students from the University of Minnesota, and graduate students in speech-language pathology at the University of Minnesota. This allowed us to examine whether the influence of visual biasing on speech perception is

mediated by experience perceiving children's speech. Our predictions here were somewhat complex. First, we predicted that less-experienced listeners overall would have a stronger influence of presumed talker race on judgments of accuracy than would more-experienced listeners. Second, we predicted that more-experienced listeners would be biased visually only for the stimuli with final /t/, as this is the variable that actually differs between speakers of African American English and standard American English. More-experienced listeners would assume that the deleted /t/ tokens were accurate when paired with photos of African American faces, and inaccurate when paired with photos of Caucasian faces. We predicted that the less-experienced listeners might be biased by the presumed race of the speaker for both the /s/-initial and /t/-final stimuli equally. However, based upon Staum Casasanto's (2008) findings wherein naïve listeners displayed implicit knowledge of African American English when presented with final consonant clusters and photos of Caucasian and African American adults, we would not have been surprised to find no differences between the two groups.

Third, this study is innovative in that we examined implicit attitudes toward race, and their influence on judgments of accuracy. Here were interested in what correlates, if any, existed between listeners' ratings and their attitudes about race. For this we used a measure of implicit bias called the Implicit Association Task (IAT) (Greenwald, McGee & Schwartz, 1998), explained in more detail in section 2. This reaction-time task yields a summary measure that has been shown in previous research to reliably index individuals' attitudes toward different social groups. Our IAT examined attitudes toward race. We predicted that we would see participants with a greater implicit pro-white bias would rate

tokens paired with a photo of a Caucasian child as more accurate than those when paired with a photo of an African American child. Conversely, we expected to see less discrepancy between ratings of tokens paired with photos of different races in participants with neutral attitudes. Because we measured both experience and attitudes, we were able to examine whether the two interacted, and specifically whether listeners with more experience perceiving children's speech had a smaller effect of implicit attitudes on the perception of the accuracy of speech paired with African American and Caucasian children's faces.

2. Methods

Two tasks were used in this project, one in which listeners judged the accuracy of children's speech, and one in which we measured listeners' implicit attitudes toward race. Briefly, the first task employed pictures of African American and Caucasian children to imply the race of the talker. These pictures were randomly paired with single word speech tokens. Some of the speech tokens were altered by replacing produced phonemes with synthesized phonemes to suggest a speech error. We then asked listeners to judge the accuracy of the speech token they heard using a visual analog scale (VAS). For the second task, we measured the listener's implicit attitudes about race using an Implicit Association Task (Greenwald et al. 1998). Each of these tasks is now described in greater detail in turn.

Speech Accuracy Judgment Task

Participants

Thirteen Speech-Language Pathology graduate students and 26 undergraduate students from the University of Minnesota participated in this study. Twenty-nine of the listeners were female and 9 were male. All participants were between 18 and 50 years of age and were native English speakers with no history of communication or hearing disorders. Listeners were recruited by posting fliers around campus and making announcements in several classes within the department: two large undergraduate introductory courses (in which the majority of those enrolled are not Speech-Language-Hearing Science majors) and four graduate courses. No mention of race or gender was made in the recruitment materials, but rather the study was presented solely as a speech

perception study. Subjects were compensated $10 for their time. Subjects completed a post-experiment questionnaire in which they rated the amount of time they spend with children, on a 10-point scale, with 1 indicating little or no experience, and 10 indicating extremely frequent experience. The average age of the undergraduate group was 23.3 year (SD = 7.3 years), while that of the graduate students was 23.4 years (SD = 3.8 years). This difference was not statistically significant. The average self-rated experience of the undergraduates was 2 (IQR = 2), while that of the graduate students was 4 (IQR = 2). A Mann-Whitney U Test examined group differences in self-ratings of experience with children. The group differences were significant (Mann-Whitney U = 92.5, Wilcoxon W = 443.5, z = -2.324, p = 0.021).

Stimuli

We manipulated single word tokens recorded for a previously conducted study to create the auditory stimuli for this experiment. In the previous study, Crocker and Munson (2006) elicited speech samples from thirty-two boys being seen as patients at the Centre for Addiction and Mental Health (CAMH) in Toronto, Ontario, to explore the development of sociophonetic variants that convey gender identity in preadolescent children. The talkers were all native speakers of English and had no history of communication or hearing disorders. They spoke the English dialect common to southern Ontario, which is spoken in Canadian national broadcast media and similar to the Midwestern American English dialect.

Of the 14 target words that were recorded by Crocker and Munson, 7 were chosen for this experiment. The original target words were common objects selected because they were easy to represent with a picture card and represented high frequency

vocabulary. All of the words used in this experiment were monosyllabic and 6 of the 7 consisted of a CVC structure. The 7 words included 2 words with word-initial /s/, 3 words that contained word-final /t/, and two filler words. (See Table 2.1 below.)

Tokens from thirty-two different talkers were used. All of the talkers were boys. However, as shown by Crocker and Munson (2006), naïve listeners perceived these boys to vary greatly in how gender typical they sounded, with some boys sounding very typically boy-like, and others sounding relatively girl-like. Productions were elicited using a picture-naming task. A flip-book consisting of one picture per page was used to elicit each target word five separate times over the course of the task. This ensured at least one fluent and accurate recording of each word. Talkers were instructed to only say the name of the object in the picture with no introductory phrases. If the talker produced a synonym rather than the target word, a prompt was given to elicit the desired token. The task was recorded in a standard consultation room in the CAMH building. An AKG C420 head-mounted micro-mic attached through a Rolls phantom power source to a Marantz CDR 330 CD recorder was fitted to each talker. A 44.1. kHz sampling rate with 16-bit quantization was used during recording and processed through a low-pass filter with an upper cutoff of 22.05 kHz to prevent aliasing.

The speech data were converted into .wav audio files in Minneapolis. Individual tokens were then extracted using the Praat signal-processing system, version 4.3.27 (Boersma & Weenink 2005). One token of each target word by each speaker was chosen at random, then discarded if the recording was deemed noisy or of poor sound quality, at

which point a new random token was chosen. The final tokens were then normalized for loudness by adjusting peak amplitudes of each token to be equal to the others.

We selected 7 word tokens for each talker to be used in this experiment. Five of these words were target words, that is, words that we manipulated acoustically and which were the target of our analysis. The other two words were filler items. All seven words are shown in Table 2.1, below. This table divides the stimuli according to whether they were in the set of words with an initial sibilant fricative, the set of words containing a final /t/, or filler stimuli. As the word tokens were produced by Canadian speakers, we intentionally chose words that did not include markedly Canadian pronunciation variants. For instance, we eliminated any words with /ɑu/ or /ɑɪ/ diphthongs as these diphthongs have Canadian variants that differ significantly from Standard American English variants.

Table 2.1. Single word tokens used in the Speech Accuracy Rating Task

Orthography	IPA	Independent variable
bee	[bi]	filler
cake	[keɪk]	filler
boot	[but]	final /t/
foot	[fʊt]	final /t/
hat	[hæt]	final /t/
sock	[sɑk]	initial /s/
sun	[sʌn]	initial /s/

To create the speech accuracy judgment task, the two initial /s/ tokens and the three final /t/ tokens of the target words were acoustically manipulated using the Praat signal processing system, version 5.1.23 on a Macintosh computer. The onset and the offset boundaries of the word-initial sibilants in [sɑk] and [sʌn] and word-final stops in [but], [fʊt] and [hæt] were manually marked using the Text Grid feature in Praat. Next a batch program replaced the [s] phonemes with a synthesized [θ] or a synthesized [s], and gated (i.e. truncated) the final stops at the burst onset boundary and at 20 ms before the burst onset boundary. This resulted in two variants for each /s/-initial word (one with a correct /s/ and one with a [θ]-for /s/ substitution), and three variants for each /t/-final

word (the whole word, gated at the onset of the final /t/ burst, and 20 ms prior to the closure for the final /t/). Recall that we chose to gate the final /t/ tokens at two different points in the stop phoneme. The variant gated at the onset of the burst simulates an unreleased final /t/; the variant gated 20 ms before the onset of the burst simulates a glottalized final /t/. The glottalized variant represents what a listener might expect to hear if listening to a speaker of African American English. As Labov (1972, p. 19) reported, the final /t/ in "boot" is glottalized in African American English, and is pronounced /buʔ/ whereas in Standard American English "boot" is pronounced /but^{h}/ or /but$^{\neg}$/.

Finally, we assigned 16 of the talkers to be represented by photos of boys and the other 16 to be represented by photos of girls. Though perceived gender was not the primary purpose of this study, picture gender was used as a factor in all of the analyses, as this variable has been shown previously to affect the perception of speech and language (i.e., Munson & Seppanen, 2009; Strand & Johnson, 1999). The decision to include pictures of boys and pictures of girls was made partly for pragmatic reasons, and partly for theoretical ones. Pragmatically, the public-access corpus of pictures that we accessed for this project had only a small number of pictures that met our criteria for use in this experiment (i.e., being sufficiently clear, having a clear ethnicity). Using both pictures of boys and pictures of girls doubled the number of usable pictures. Second, the corpus from which these stimuli were drawn included boys with a range of gender-typicality of voice to mitigate the concern that pairing a less boy-like sounding talker with a picture of a boy would affect the audiovisual integration of talker race. Third,

including pictures of boys and girls increased this study's ecological validity. If the study included only pictures of boys, then it would limit how reflective it was of real-world situations in which listeners interact with children.

Pictures and talkers were matched by the author of this thesis by listening to tokens by each talker and assigning gender categories subjectively. This resulted in a corpus of 224 original and accurately produced speech tokens plus 160 speech tokens which were altered to sound as if they were inaccurately produced, for a total of 384 tokens from 32 talkers who were assigned gender designations.

In order to bias listeners to impute racial categories onto talkers, tokens were randomly paired with a picture of an elementary age girl or boy who appeared to be either Cacuasian or African American. Eight photos for each category – Caucasian girls, Caucasian boys, African American girls, African American boys – were used. We retrieved photos from the internet as well as from a CD of stock photos that was purchased by the University of Minnesota, Twin Cities (Eyewire Images, 2002). The photos were resized in Photoshop to 4" X 6" dimension. Practice items were also included. For the practice items, we used 5 photos of children who appeared to be of Asian ethnicity (3 girls and 2 boys). We also used speech tokens that were not included in the main experiment.

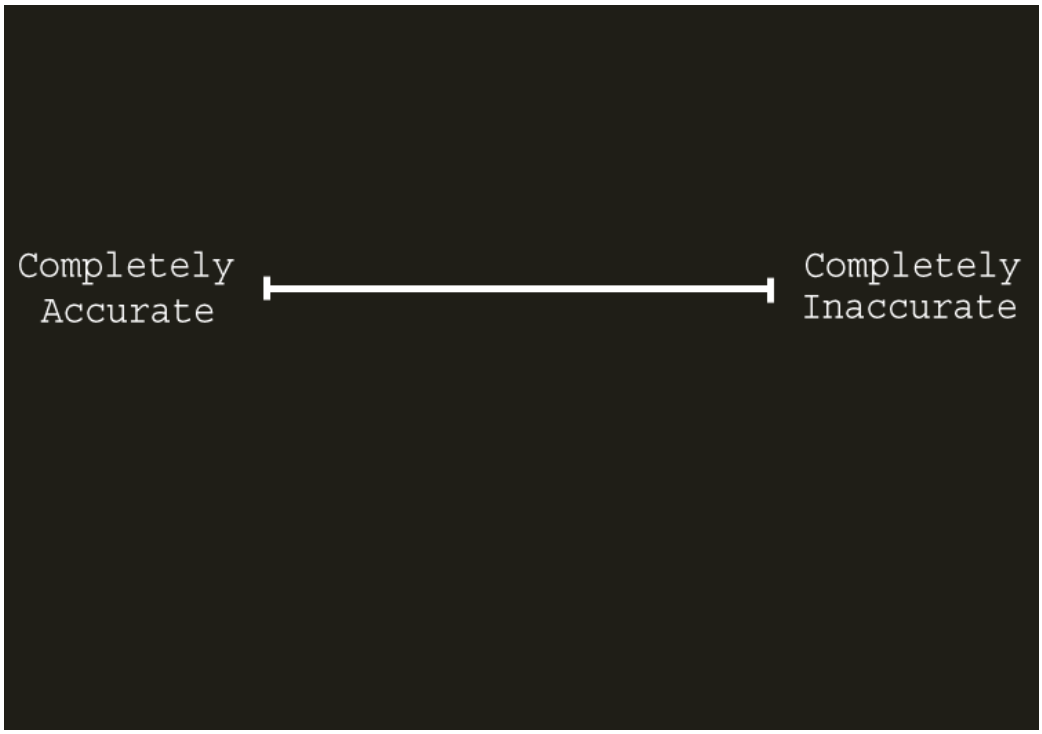
We used a random order generation program to assign one of the three word-final stop conditions (unaltered, gated at the burst, or gated 20 ms before the burst) for each target word and one of the 2 sibilant conditions (/s/ or /θ/) for each target word then randomly pairing each talker's tokens that were assigned a female or male designation by

the thesis author with a photo of either a Caucasian girl or boy or with a photo of a African American girl or boy respectively. Each listener judged the accuracy of 224 tokens, either an original or an altered version of each target word from each talker. Each talker was consistently represented by the same photo within each experimental condition, but was purposefully different between conditions.

Procedure

The speech accuracy judgment task was programmed and run using the E-prime experiment management software (Schneider, Eschman, & Zuccolotto, 2002. Version 1.2). Listeners sat at a computer monitor placed approximately at eye-level in a sound-treated booth. Instructions for the task were provided on the computer monitor. The speech tokens were peak normalized and delivered via headphones at approximately 70 dB SPL. Listeners were given 5 practice trials immediately preceding delivery of the target tokens. On each trial, listeners first viewed a photo of a child whom they had been instructed at the beginning of the task to imagine as the talker. Then, they listened to a token via headphones while the screen flashed an orthographic transcription of the target word in 36-point courier font. Next the listener was asked to rate the accuracy of the speech sample using a visual analog scale (VAS). The VAS consisted of a horizontal line with two endpoints labeled with “Completely Accurate” and “Completely Inaccurate” (illustrated in Figure 2.1 below).

Figure 2.1. A screenshot of the Visual Analog Scale. The horizontal line on the screen began at 90 pixels from the left hand side of the screen and ended at 535 pixels from the left hand side of the screen.



The listeners were instructed to point and click with a mouse the place on the line which approximated where they felt the given token fell along the Accurate-to-Inaccurate continuum. The distance along the line at which they clicked (in pixels, ranging from 90 pixels at the "Completely Accurate" end and 535 pixels at the "Completely Inaccurate" end) was recorded and later analyzed.

Implicit Association Task

The Implicit Association Task (IAT) we used was created by Babel (2009) for a previous experiment that also examined the correlation between implicit biases about

race and the effects of perceived race on speech perception. All listeners completed the IAT immediately following the speech accuracy judgment task

The stimuli for the IAT consisted of 20 stereotypically African American names (associated with 'BLACK') and 20 stereotypically Caucasian names (associated with 'WHITE') and 20 words that have positive associations (associated with 'good') and 20 words that have negative associations (associated with 'bad'). They are listed in Table 2.2 below. The names and words used by Babel were taken from those used in Greenwald et al. (1998), Dasgupta and Greenwald (2001), and Jelenec and Steffens (2002). No control for familiarity with the names was considered as Dasgupta et al. (2000) found that familiarity in this IAT design had no significant affect.

Table 2.2. Stimuli used in the Implicit Association Task. Names or words in each column received a CORRECT response only if the respondent chose the corresponding association category (BLACK, WHITE, good, or bad).

Stereotypically African American ('BLACK') names	Stereotypically Caucasian American ('WHITE') names	Words associated with 'good'	Words associated with 'bad'
Aaliyah	Abby	caress	abuse
Aijia	Amy	cheer	agony
Alonzo	Carson	diamond	awful
Andre	Claire	freedom	cancer
Dominique	Cody	friend	crash
Ebony	Colin	gentle	death
Jada	Connor	glorious	evil
Jamal	Dustin	happy	failure
Jazmine	Emily	health	filth
Latonya	Hannah	honest	horrible
Marquis	Heather	joy	hurt
Maurice	Jack	laughter	jail
Raven	Jake	love	murder
Shanice	Jenna	loyal	nasty
Temeka	Katherine	lucky	poverty
Terrance	Katie	paradise	rotten
Terrell	Logan	peace	sickness
Tiara	Luke	pleasure	terrible
Trevor	Madeline	rainbow	tragedy
Tyrone	Scott	wonderful	vomit

Procedure

The task was administered at the same computer monitor as the previous task using a 5-point equal interval button-box to record responses. Instructions were provided both on the computer screen and by a laboratory assistant. The buttons were labeled 1 to 5, 1 being the left-most button. Listeners were told to use their dominant hand to select the button labeled with a "1" to correspond with the category displayed on the left side of the computer screen, or the button labeled with a "5" to correspond with the category displayed on the right side of the screen.

The task consisted of 5 blocks: target-concept discrimination, associated attribute discrimination, combined test, target-concept discrimination (concepts reversed), and reversed combined task. For the first block, the target categories BLACK and WHITE were presented in the upper corners of the computer screen opposite each other. Individual names were then presented in the center of the screen. The listeners' task was to categorize the name on the screen as BLACK or WHITE as quickly and as accurately as possible using the button-box. The second block replaces BLACK and WHITE with 'good' and 'bad' so that listeners categorize the words semantically. The third block combined the first two blocks so that the attribute ('good' or 'bad') appeared under the concept (BLACK or WHITE). Randomly selected names (in all capital letters) and words (in lower case) were presented to the listeners who were asked to categorize them using the button box, but ignoring the attribute for names and ignoring the concept for words. The fourth block reversed which side of the screen BLACK and WHITE appeared on from the first block, but otherwise was the same. The fifth block reverses

the third block, so that if 'good' appeared under WHITE for the third block, it would appear under BLACK for the fifth block. Figure 2.2 below illustrates an example of the third and fifth blocks.

Figure 2.2. An example of the screen presented to respondents during the Implicit Association Task in the third or fifth blocks.

We calculated listeners' IAT scores using the methods described in Greenwald, Nasek and Banaji (2003). In this scoring system, negative IAT scores indicate a pro-Black bias, positive a pro-White bias, and zero no bias. The average IAT of the undergraduate group was 0.56 (SD = 0.44), while that of the graduate students was 0.60 (SD = 0.37). This difference was not significant. IAT scores did not correlate with self-reported experience perceiving children's speech.

Debriefing

At the end of the experiment, after participants completed both tasks, we conducted a debriefing in which we explained that the primary focus of the study was race (which we purposefully withheld until that point.) Participants were offered the opportunity to withdraw their data if they wished. No one chose to withdraw. A handout with details about the object of the study and information of how to withdraw from the study if desired was given to each participant for his or her records. A copy of the handout can be found in Appendix A.

3. Analysis

Two statistical models were used to analyze the results of the speech accuracy rating task. First, these data were subjected to two analyses of variance, one for the ratings of fricative stimuli and one for the stop stimuli. Both sets of ratings were then subjected to separate linear mixed effects analyses. The latter analysis was used to examine the influence of IAT scores and self-reported experience on speech-accuracy ratings.

ANOVA

Before running the analyses of variance on the speech accuracy rating task data, average accuracy ratings were calculated for each of the 39 listeners as follows. Responses to each unaltered and manipulated word token category (Fricative Type: simulated word-initial /s/ and substituted word-initial /θ/; Stop Type: word final /t/ unaltered, gated at the burst, and gated 20 ms before the burst) were each grouped by the race and gender depicted in the photograph they were paired with (African American girl, African American boy, Caucasian girl, and Caucasian boy), then averaged, for a total of twenty ratings per subject (five manipulation conditions, by 2 picture genders, by 2 picture races). The averaged ratings were the dependent measures in two separate analyses of variance (ANOVA): one on Fricative Type data and one on Stop Type data. The independent variables were Listener Group (undergraduate student or Speech-Language Pathology graduate student), Photograph Gender (male or female), and Photograph Race (African American or Caucasian). The accuracy ratings used in both analyses were determined by a click on the x-axis of a Visual Analog Scale: a horizontal

line bounded by “Completely Accurate” on the left and “Completely Inaccurate” on the right which ranged from 90 to 535 pixels, as described in the Methods section, above. A click on the end point on the “Completely Accurate” end of the scale was logged as 90, and a click on the end point on the “Completely Inaccurate” end of the scale was logged as 535.

Linear Mixed Effects Analysis (LME)

Two linear mixed effects analyses were performed on individual accuracy ratings: one on Fricative Type data and one on Stop Type data. The LME model combines two types of independent measures in the analysis: random effects and fixed effects (Baayen, Davidson, & Bates, 2008). In this study the random effects analyzed were participant and item. Items were comprised of every possible permutation of target word token and photograph pairings. The fixed effects common to these analyses were Photograph Gender, Photograph Race and two factors not included in the ANOVA: Child Time (the experience in years each participant reported having with children) and IAT score. In the Stop Type analysis, an additional fixed effect was stimulus type. The dependent measure consisted of each independently made accuracy rating elicited by each item (i.e., each unique token-photograph pairing) as indexed by the click location on x-axis of the Visual Analog Scale. The benefit of the LME model over ANOVA is that it analyzes each item separately rather than averaging items. Unlike ANOVA, the LME model does not assume that each item will 'behave' the same way, as is assumed when groups of stimuli are averaged together. Moreover, LME, unlike ANOVA, allows for the examination of how different participant characteristics (here, IAT score and self-reported experience

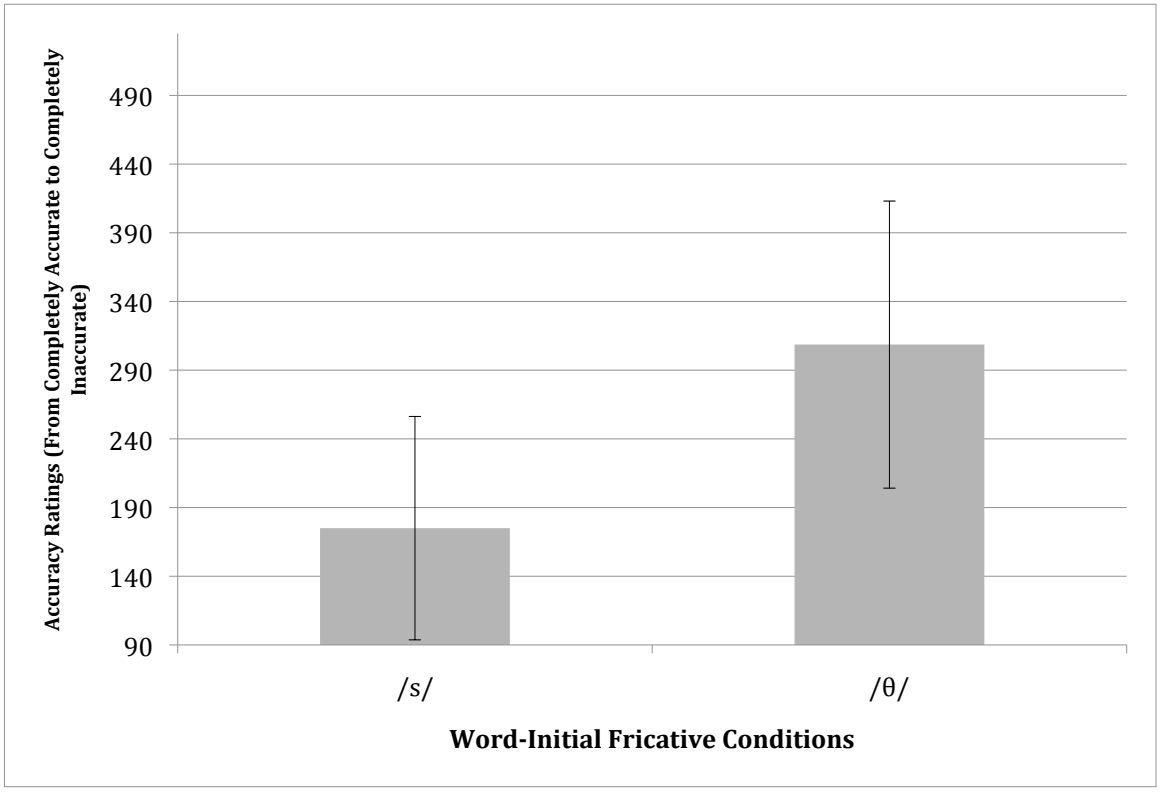
perceiving children's speech) affect a dependent measure (here, accuracy ratings), and how they interact with other fixed effects (here, picture race and picture gender).

4. Results

ANOVA on Fricative Type Data

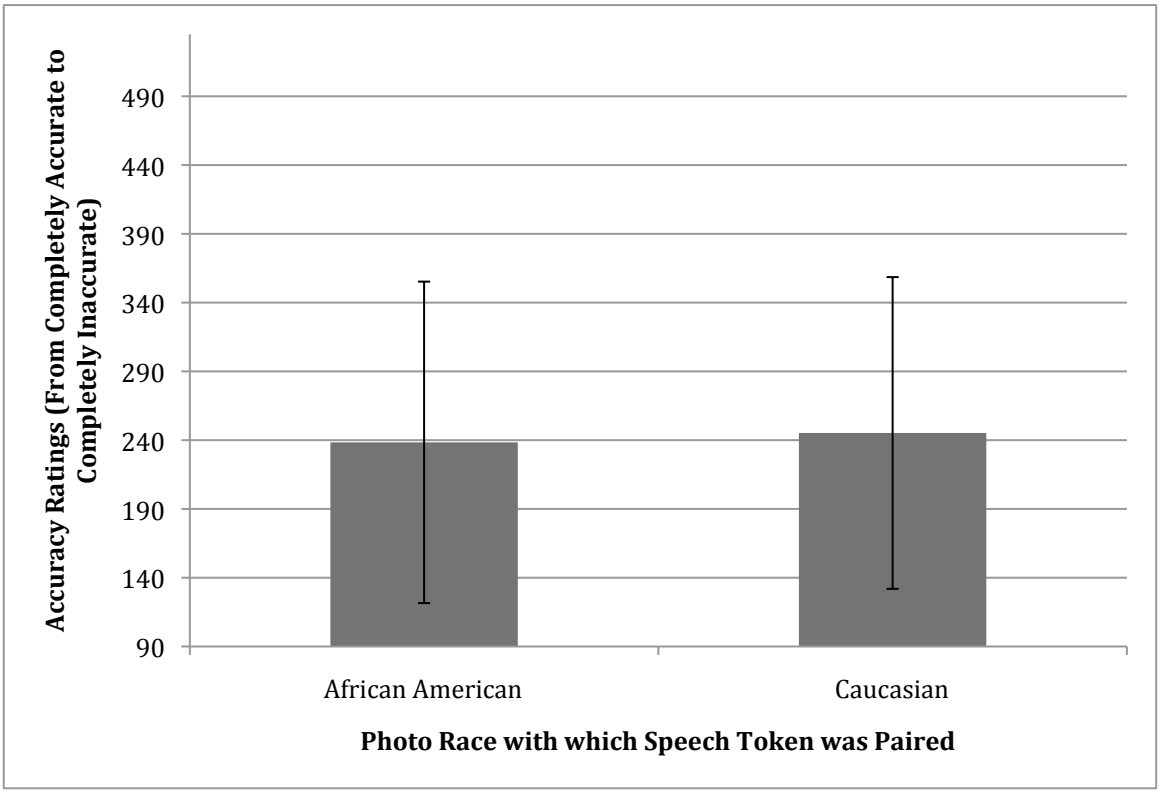
The analysis of variance on Fricative Type data revealed two main effects and one interaction. Recall that averaged Accuracy Ratings were the dependent variable, while the Fricative Type, Photograph Gender, and Photograph Race were the independent variables. A significant main effect of medium size (using the guidelines for interpreting effect sizes presented in Cohen, 1988) was observed for fricative type ($F[1,37] = 43.515$, $p<0.001$, partial $\eta^2 = 0.54$). This effect is illustrated by the differences between mean accuracy ratings in Figure 4.1, below. In viewing this figure, the reader is reminded that the higher the accuracy rating, the more inaccurate the token was deemed (i.e., closer to the "completely inaccurate" end of the visual analog scale). Participants rated the word-initial /θ/ as significantly less accurate than word-initial /s/.

Figure 4.1. Accuracy ratings of word tokens with word-initial /s/ vs. those with a substituted word-initial /θ/. (Lower accuracy rating values indicate listener judged production as more accurate.)



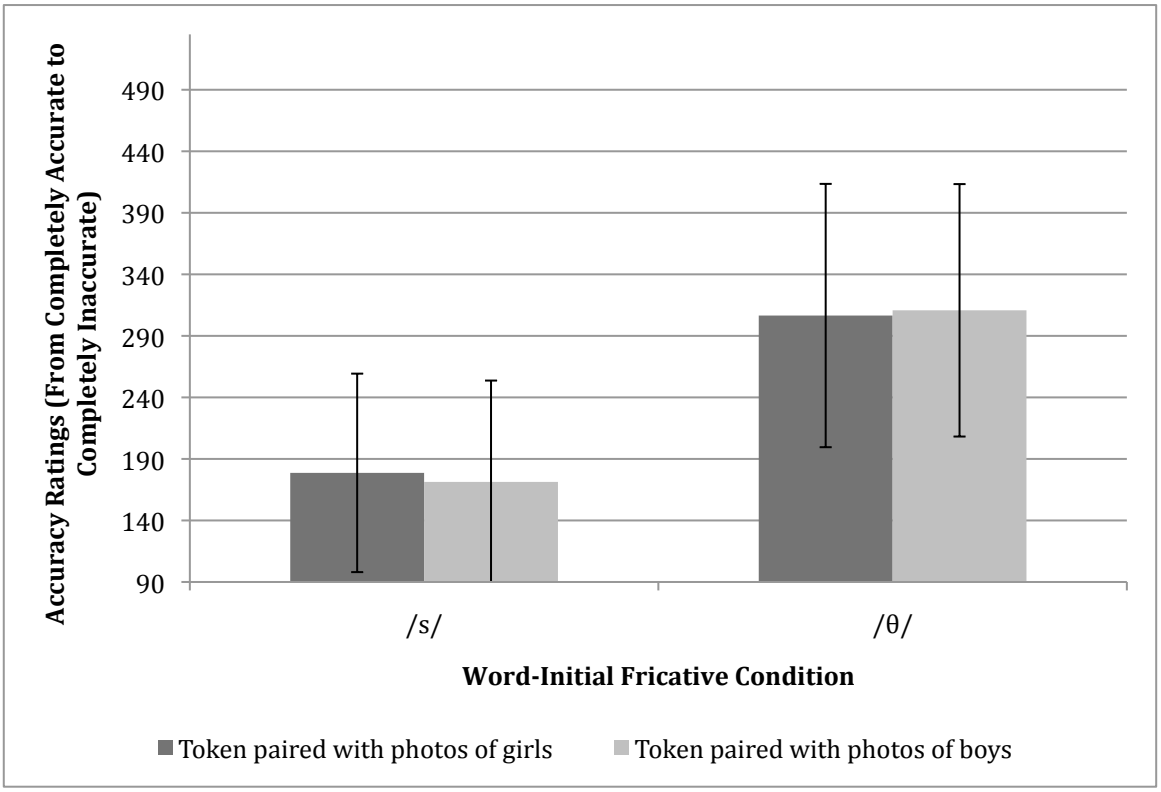
A another main effect of small size was noted for Photograph Race (F[1,37] = 4.824, p=0.034, partial η^2 = 0.115) which can be seen by comparing the bar heights in Figure 4.2, below. Tokens with word-initial /s/ and /θ/ were rated as slightly more accurate when paired with a photograph of an African American Child and slightly less accurate when paired with a photograph of a Caucasian child.

Figure 4.2. Accuracy ratings of word tokens with word-initial /s/ and /θ/ separated by Photograph Race as determined by an analysis of variance of Fricative Type data. (Lower accuracy rating values indicate listener judged production as more accurate.)



A nearly significant interaction was found between Fricative Type and Photo Gender (F[1,37] = 3.618, p=0.065, partial η^2 = 0.089) which can be seen by comparing the bar heights in Figure 4.3, below. Tokens with word-initial /s/ were rated as slightly less accurate when paired with photos of girls than when paired with photos of boys. Conversely, tokens with word-initial /θ/ were rated as slightly more accurate when paired with photos of girls than when paired with photos of boys.

Figure 4.3. Accuracy ratings of word tokens with word-initial /s/ vs. those with a substituted word-initial /θ/ divided by Photograph Gender as determined by an analysis of variance of Fricative Type data. (Lower accuracy rating values indicate listener judged production as more accurate.)



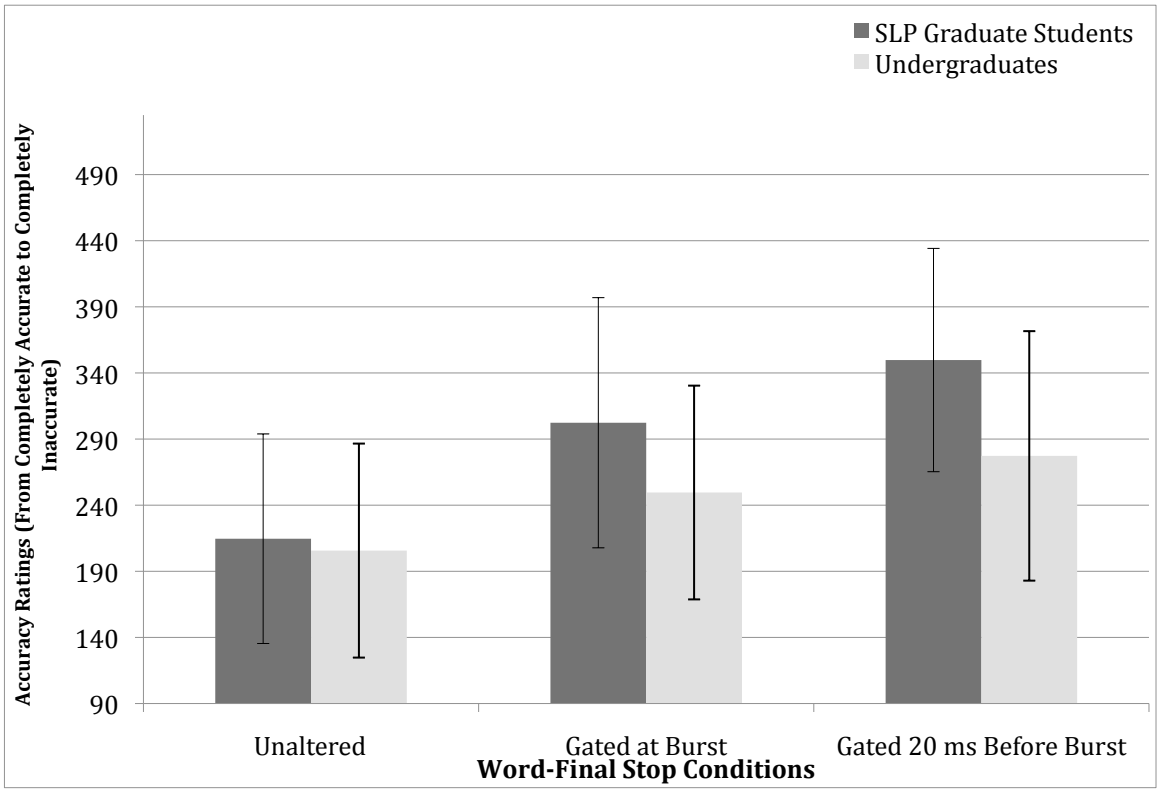
ANOVA on Stop Type Data

The analysis of variance on Stop Type data revealed several significant main effects and interactions. Averaged Accuracy Ratings were treated as the dependent variable, while the Stop Type, Photograph Gender, and Photograph Race were entered as the independent variables. A significant main effect was observed for Stop Type ($F[2,74] = 90.543$, $p<0.001$, partial $\eta^2 = 0.71$) as illustrated by comparing bar heights in

Figure 4.4. As above, the higher the accuracy rating, the more inaccurate the token was deemed. Listeners rated tokens with word-final /t/ as significantly more inaccurate as more of the final stop was gated off. Tokens that simulated an unreleased final /t/ (gated at the stop burst) were rated as more inaccurate than unaltered final /t/ tokens, and tokens that simulated a glottalized final /t/ (gated 20 ms before the burst) were rated as even more inaccurate.

A significant yet small interaction between stop type and whether or not the listener was a Speech Language Pathology (SLP) graduate student or an undergraduate was evident ($F[2,74] = 8.737$, $p<0.001$, partial $\eta^2 = 0.191$) and is also illustrated by Figure 4.4 below. The mean accuracy ratings increased as a larger portion of the final stop was gated off for both groups. However, ratings of inaccuracy increased at a significantly greater rate for graduate students than for undergraduates. While both groups rated the speech tokens as more inaccurately produced the more the token was gated, the graduate students perceived the tokens in which the final stop was most altered to be less accurate than the undergraduate students did.

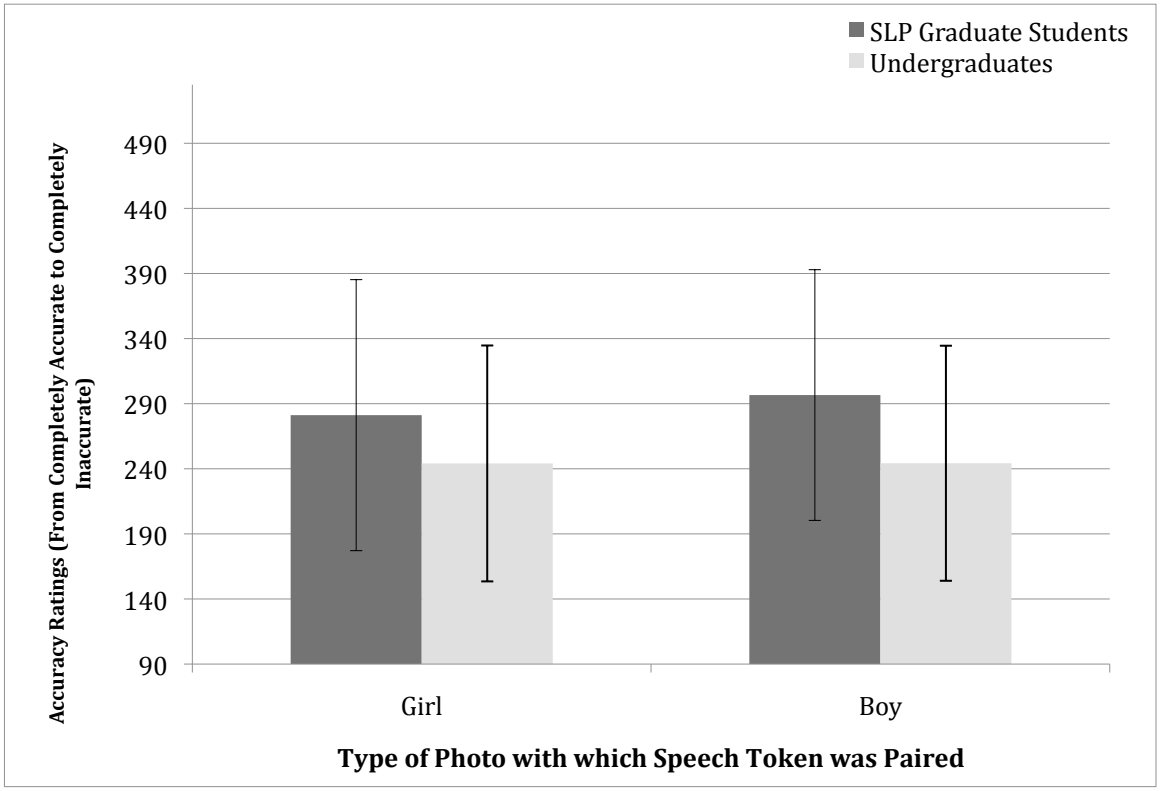
Figure 4.4. Accuracy Ratings by SLP Graduate Students and Undergraduates of Speech Tokens with Word-Final /t/ Unaltered, Gated at the Burst, and Gated 20 ms Before the Burst. (Lower accuracy rating values indicate listener judged production as more accurate.)



Another significant main effect found by ANOVA with respect to stop type was evident for perceived gender (F[1,37] = 5.724, p=0.022, partial η^2 = 0.134). Tokens with word-final stops that were paired with photographs of girls were rated as slightly more accurate than when paired with photographs of boys (see Figure 4.5, below). A significant yet weak interaction was noted between perceived gender and listener group (graduate student or undergraduate) (F[1,37] = 5.520, p=0.024, partial η^2 = 0.130) which can be

seen by comparing the bar heights in Figure 4.5, below. While undergraduates rated the accuracy of all speech tokens as virtually identical regardless of the gender of the child in the photograph with which the tokens were paired, graduate students in Speech-Language Pathology rated speech tokens paired with photographs of boys as slightly less accurate than those paired with photographs of girls.

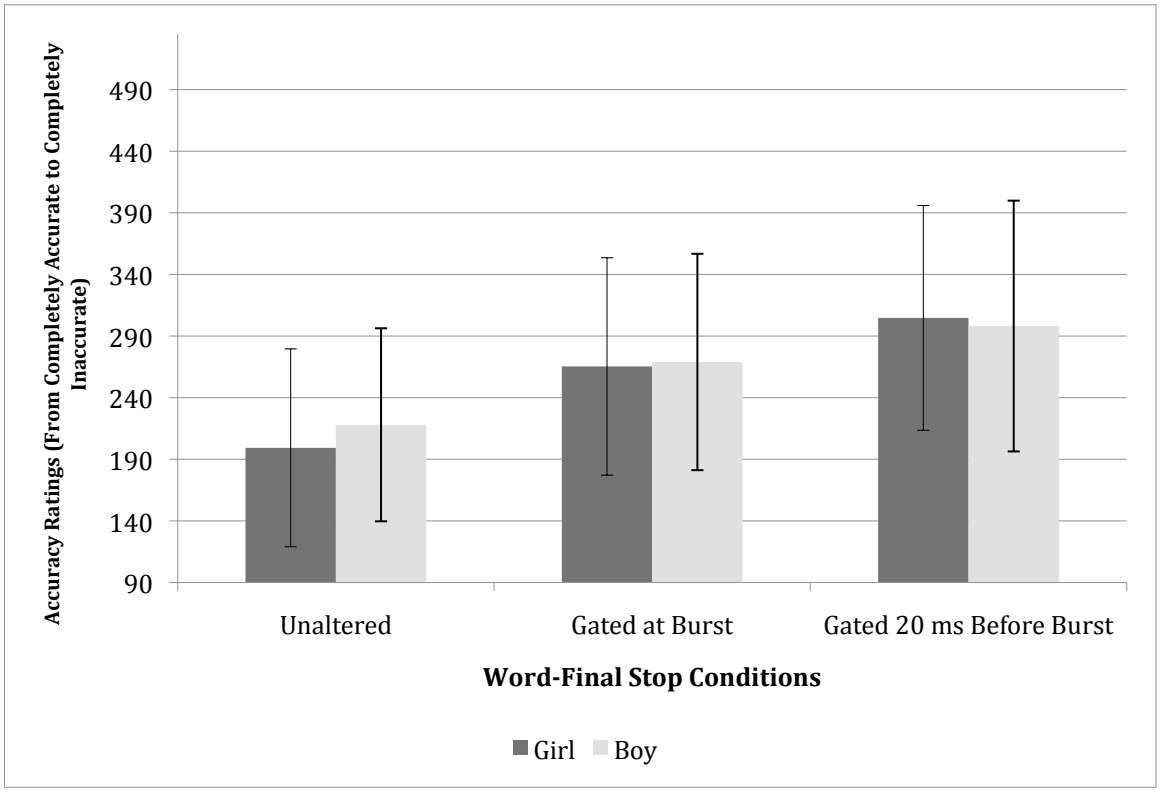
Figure 4.5. Accuracy Ratings Made by SLP Graduate Students and Undergraduates of Speech Tokens Paired with a Photograph of a Girl or a Boy. (Lower accuracy rating values indicate listener judged production as more accurate.)



A significant yet weak interaction was also present between photograph gender and stop type ($F[2,74] = 6.217$, $p=0.003$, partial $\eta^2 = 0.144$) seen in Figure 4.6.

Unaltered speech tokens with a word-final /t/ were rated as slightly more accurate by both groups of listeners when paired with a photograph of a girl than when paired with a photograph of a boy.

Figure 4.6. Accuracy Ratings of Speech Tokens with Word-Final /t/ Unaltered, Gated at the Burst, and Gated 20 ms Before the Burst When Paired with a Photograph of a Girl or a Boy. (Lower accuracy ratings indicate listener judged productions as more accurate.)



Linear Mixed Effects Analysis (LME)

The linear mixed effects model found interactions between two of the independent variables which we examined in the ANOVA, photograph race and photograph sex, and two other measures we gathered: the Implicit Association Task score (IAT) and a self-report of the exposure participants had to children's speech reported as

time spent with children in years (Child Time). Group (undergraduate versus graduate student) was not included in this analysis.

Two separate LME analyses were conducted: one on Fricative Type data and the other on Stop Type data. Table 4.1 presents the data for the significant effects and interactions for each analysis.

Table 4.1. Significant effects and interactions reported by a mixed effects linear model.

Effect	β Estimate	Standard Error	*t*-value	*p*-value
LME – Fricative Type Data				
Intercept	172.274	46.332	3.718	<0.001
Fricative Type-θ	113.905	29.137	3.909	<0.001
IAT : Fricative-θ	94.873	37.743	2.514	<0.001
LME – Stop Type Data				
Intercept:	173.565	49.312	3.520	<0.001
IAT Score	128.176	63.804	2.009	0.0160
Child Time	33.721	13.507	2.496	0.0126
Race : Gender : Stop-Unaltered	-164.362	57.872	-2.840	0.0044
Race : Child Time : Stop-Unaltered	-22.5111	11.211	-2.006	0.0442
Race : Gender : Child Time : Stop-Unaltered	33.710	15.885	2.122	0.0354
IAT : Race	75.207	37.411	2.010	0.0404
IAT : Race : Gender	-109.005	52.989	-2.057	0.0366
IAT : Race : Gender : Stop-Unaltered	181.308	75.803	2.392	0.0168
IAT : Race : Gender : Stop-Gated at (Burst-20 ms)	143.212	72.413	1.978	0.0446
IAT : Race : Child Time : Stop-Unaltered	31.165	15.965	1.952	0.0490
IAT : Race : Child Time : Stop-Gated at (Burst-20 ms)	27.861	14.578	1.911	0.0524

In interpreting this table, the reader is reminded that LME models predict outcomes from fixed and random residual factors of individual items (i.e., there is no

averaging across subjects or items). The LME takes nominal variables (i.e., fricative type, which has two levels, /θ/ and /s/, and stop-type, which has three levels) and recodes them as contrasts between one variable and the other variable (or, in the case of stop-type, the other two variables). For example, when the table refers to the 'Unaltered Stop' condition it is actually referencing the contrast between the Unaltered Stop condition and the other two.

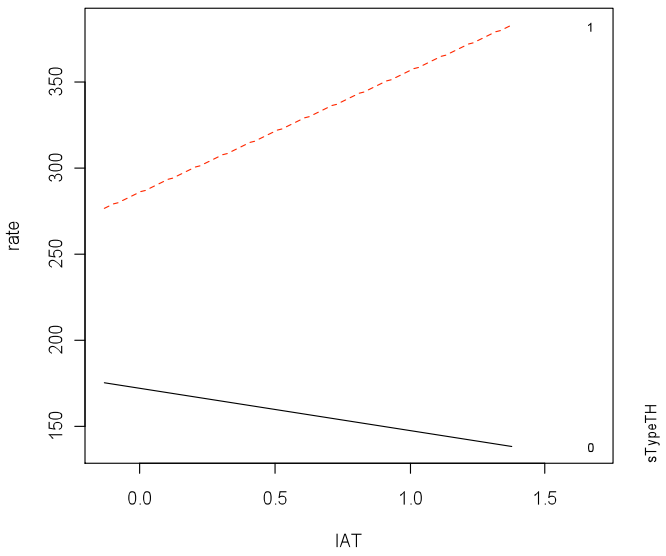
This next section covers each significant finding for each of the two analyses. We have included detailed interpretations of lower-order effects even when they were mediated by higher-order interactions in the interest of helping the reader best understand the complex set of interactions among the factors that are at play in each very complex analysis. This analysis is supplemented with numerous figures. These figures differ from the earlier figures in this section in that they represent model predictions, rather than observed data, as is conventional when representing the results of LME (i.e. Baayen et al., 2008). The predicted rating is a function of the intercept (which says what the rating would be if the other independent variables were zero) and the slope of the line. The intercepts and the slopes vary in these LMEs, hence, attention must be paid to both slopes and intercepts when comparing the sets of graphs that represent either a main effect or an interaction. The y-axis in every graph represents accuracy ratings. As in the previous figures, lower values indicate a rating of greater accuracy. The x-axis varies from figure to figure, and represents either participants' experience level with children (Child Time) or Implicit Association Score (IAT). Recall that a score of zero on the IAT indicates a neutral attitude towards race (white/black), a positive score indicates a pro-white bias in

attitude, and a negative score indicates a pro-black bias in attitude. The functions on the graph indicate the race of the photograph the auditory stimuli were paired with when being rated (AA for African American and C for Caucasian). Whenever gender is part of the interaction, graphs on the left side of the page represent ratings of stimuli paired with girls' photographs, and graphs on the right side of the page represent stimuli paired with boys' photographs.

Linear Mixed Effects Analysis (LME) of Fricative Type

The LME of Fricative Type reported a main effect for Fricative Type [β=113.905, $t(3925)=3.909$, $p < 0.001$). The positive values associated with this factor shows that listeners as a group rated the stimuli with /θ/ as less-accurate than those with /s/. This is consistent with the results of the ANOVA. The LME of Fricative Type also reported a two-way interaction between IAT Score, by Fricative Type [β=33.710, $t(3925)=2.122$, p=0.0345] as seen below in Figure 4.7. The positive slope associated with this interaction term shows that people with higher IAT scores (i.e., greater pro-white bias) exhibited a bigger difference in accuracy ratings between tokens with /s/ and tokens with /θ/ than did people with lower IAT scores (i.e., more neutral attitudes).

Figure 4.7. Interaction between accuracy ratings of word tokens with word-initial /s/ (solid line) vs. those with a substituted word-initial /θ/ (dotted line) and IAT score as determined by a LME analysis of Fricative Type data. (Lower accuracy rating values, along the y-axis, indicate listener judged production as more accurate. Lower IAT scores, along the x-axis, indicate neutral racial attitude; higher IAT scores indicate pro-white bias.)



Linear Mixed Effects Analysis (LME) of Stop Type

Numerous main effects and interactions were found in the LME analysis of the stop-type data. This section deviates from the convention of moving from a discussion of simple effects before discussing lower-order interactions before discussing higher-order interactions. Specifically, this section first presents the lower-order main effects, followed by a discussion of the interactions relating to child-time, followed by a

discussion of the interactions related to IAT, and ends with a discussion of the interactions between child-time and IAT.

The LME of Stop Type reported a main effect for IAT Score [β=128.176, t(3925)=2.009, p=0.0160]. The positive β value suggests that individuals with higher IAT scores (i.e., a greater pro-white bias) were more likely to rate words as less accurate than those with lower IAT scores (i.e., a more neutral attitude). A main effect was also found for the amount of time participants reported spending with children (referred to henceforth as Child Time) [β=33.721, t(3925)=2.496, p=0.0126]. The positive values of the coefficients show that individuals who report spending more time with children were more likely to rate words as less accurate than those who reported spending less time with children.

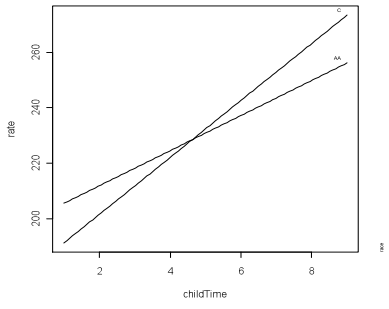
There was a significant interaction between photo gender, photo race, and condition, [β=-164.362 , t(3925)=-2.840, p=0.0044]. This effect occurred because the effect of picture race on ratings was not equivalent across picture genders and stimulus types. Specifically, there was a larger difference between ratings of tokens paired with African American girls' faces than with Caucasian girls' faces in the whole-word and -20 ms conditions than there were for the other four conditions (i.e., ratings of tokens paired with boys' faces for all three stop types, and ratings of tokens paired with girls' faces in the closure condition).

A three-way interaction was seen among Photograph Race, Child Time, and the 'Stop-Whole' factor (which indicates the contrast in ratings between the stimuli in which the word tokens were presented, and those with final stops that were unaltered) [β=-

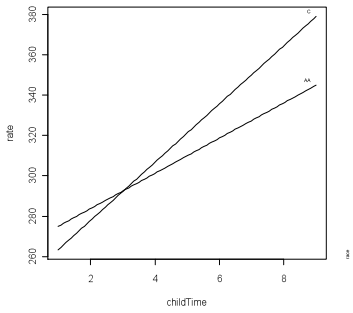
164.362, $t(3925)=-2.840$, $p=0.0044$]. Overall, as experience with children increased, so did participants' judgments of inaccuracy for tokens paired with photographs of Caucasian and African American children. However, as illustrated in the three graphs below (Figure 4.8), discrepancies between accuracy judgments of tokens paired with photographs of African American children decreased between those who spent less time with children (referred to henceforth as *Inexperienced*) and those who spent more time with children (*Experienced*) as more of the final stop was gated. Experienced participants tended to rate tokens paired with photographs of African American children as more inaccurate the more the final stop was gated, however, their judgments became closer to those of Inexperienced participants as more of the final stop was gated. The pattern of discrepancies between accuracy judgments made by Experienced and Inexperienced participants on tokens paired with photographs of Caucasian children was not unidirectional. The final stop that was gated at the burst elicited more similar judgments from Experienced and Inexperienced participants than either the unaltered tokens or the tokens gated 20 ms before the burst.

Figure 4.8. Interaction between accuracy ratings of speech tokens with word-final /t/ unaltered, gated at the burst, and gated 20 ms before the burst when paired with a photograph of an African American (AA) or Caucasian (C) child and self-reported experience with children in years (Child Time) as determined by a LME analysis of Stop Type data. (Lower accuracy rating values, along the y-axis, indicate listener judged production as more accurate. Lower Child Time values, along the x-axis, indicate less experience with children; higher Child Time values indicate more experience with children.)

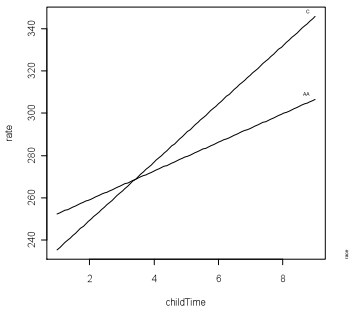
Word Final /t/ Unaltered



Word Final /t/ Gated at Burst



Word Final /t/ Gated 20 ms before Burst



The next set of graphs (Figure 4.9) illustrates a four-way interaction among Photograph Race, Photograph Gender, Child Time, and Stop Type [β=33.710, t(3925)=2.122, p=0.0345]. These graphs are separated by gender and by stop type. The data for photographs of girls are shown in the graphs on the left, and those for photographs of boys are on the right. The first set of graphs is for the unaltered final stop condition, the second set is of the tokens gated at the burst, and the third set of graphs is

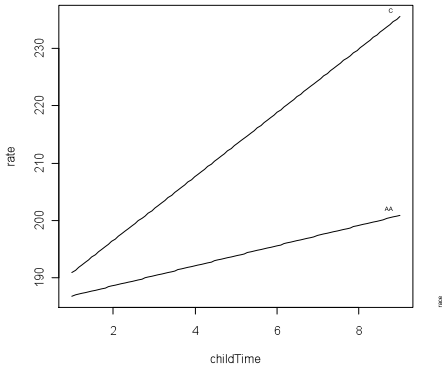
of the tokens gated at 20 ms before the burst. These illustrate that those participants who spent more time with children (experienced participants) rated tokens paired with photographs of Caucasian children as less accurate than when the tokens were paired with African American children's photographs. Interestingly, the size of the discrepancy between these ratings made by experienced participants of tokens paired with photographs of Caucasian and African Americans is significantly larger when the photograph was of a girl, and smaller when the photograph was of a boy in both the unaltered condition and the minus 20 ms condition. The greatest discrepancy was seen in the minus 20 ms condition.

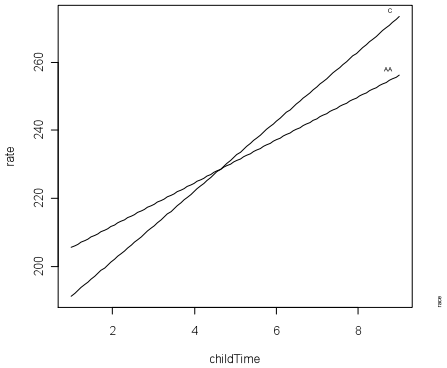
Figure 4.9. Interaction between accuracy ratings of speech tokens with word-final /t/ unaltered, gated at the burst, and gated 20 ms before the burst when paired with a photograph of an African American (line labeled AA) girl, African American boy, Caucasian (line labeled C) girl, or Caucasian boy and self-reported experience with children (Child Time) as determined by a LME analysis of Stop Type data. (Lower accuracy rating values, along the y-axis, indicate listener judged production as more accurate. Lower Child Time values, along the x-axis, indicate less experience with children; higher Child Time values indicate more experience with children.)

Tokens paired with Photos of Girls

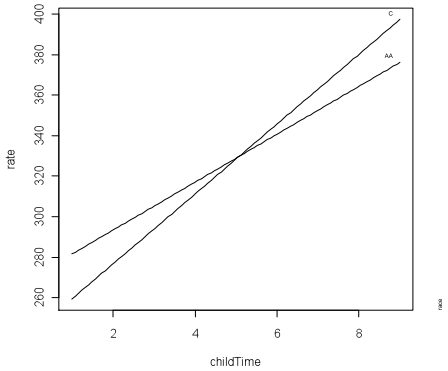
Tokens paired with Photos of Boys

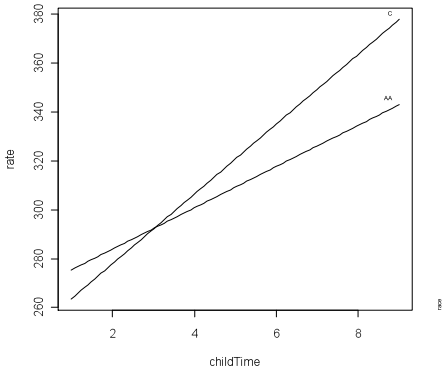
Word Final /t/ Correct



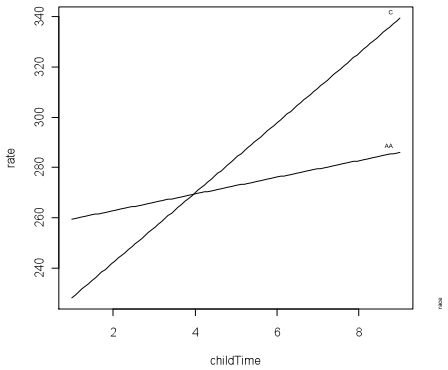


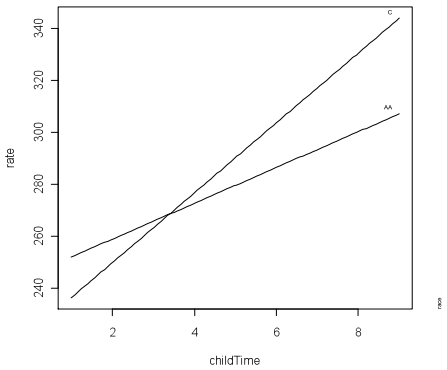
Word Final /t/ Gated at Burst





Word Final /t/ Gated 20 ms before Burst

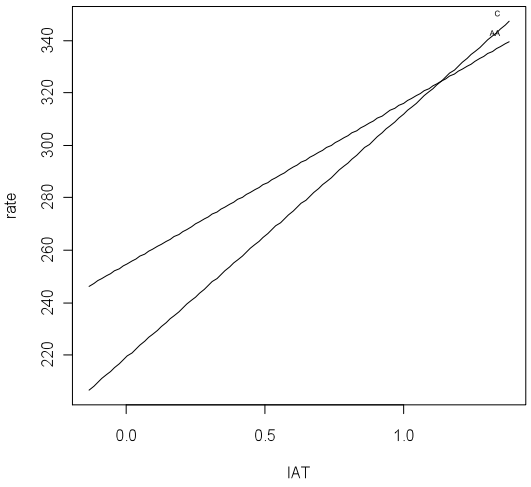




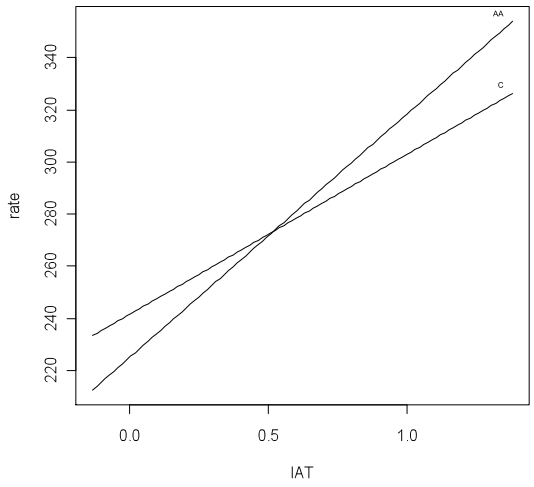
A three-way interaction between IAT Score, by Photograph Race, by Photograph Gender was reported [β=-109.005, t(3925)=2.057, p=0.0366]. As shown in the two graphs below (Figure 4.10), when the token words were appended to photographs of girls' faces (graph on the left), the interaction was such that people with greater pro-white biases rated tokens paired with Caucasian faces as slightly more inaccurate than those paired with African American faces, while those with neutral IAT scores had a tendency to rate tokens paired with African American photographs as less accurate. When paired with photographs of boys' faces, ratings mirrored the group interaction: pro-white listeners rated words as more accurate when paired with a Caucasian face, while neutral listeners rated words as more accurate when paired with an African American face.

Figure 4.10. Interaction between accuracy ratings of speech tokens when paired with a photograph of an African American (line labeled AA) girl, African American boy, Caucasian (line labeled C) girl, or Caucasian boy and Implicit Association Task score (IAT) as determined by a LME analysis of Stop Type data. (Girls' data is on the left, boys' data is on the right. Lower accuracy rating values, along the y-axis, indicate listener judged production as more accurate. Lower IAT scores, along the x-axis, indicate neutral racial attitude; higher IAT scores indicate pro-white bias.)

Tokens paired with Photos of Girls



Tokens paired with Photos of Boys



Four-way interactions took place between IAT Score, by Race, by Gender, by Stop Type-Whole [β=181.308, t(3925)=2.392, p=0.0168] and IAT Score, by Race, by Gender, by Stop Type-20 ms [β=143.212, t(3925)=1.978, p=0.0446]. Figure 4.11 consists of graphs of the ratings (on the Y-axis) by IAT Scores (on the X-axis), separated by Photograph

Gender and Stop Type. The data representing tokens paired with photographs of girls are on the left, and those paired with photographs of boys are on the right. The first set of graphs is of the unaltered token condition, the second set is of the gated at the burst condition, and the third is of the minus 20 ms condition.

This very complex interaction can be summarized as follows. For those tokens paired with photographs of boys' faces, the interaction only occurred in the gated at the burst condition. However with those tokens paired with photographs of girls the interaction was always present and complex. The magnitude of the discrepancy between photographs of African American faces and Caucasian faces is not equivalent across IAT scores for the three conditions. When the token is unaltered, as pro-white bias increases, accuracy ratings for tokens paired with a Caucasian face increase as well, while ratings for tokens paired with African American faces decreases. For both altered Stop Types, as pro-white bias increases, so do ratings of inaccuracy overall. In the closure condition, however, there is a larger discrepancy between ratings for those with a more neutral attitude tend where tokens paired with African American faces are rated as more accurate than those paired with Caucasian faces. The ratings for those with a greater pro-white bias essentially converge. In the minus 20 ms condition, more neutral participants' ratings are almost the same regardless of the race of the photograph and more pro-white participants' ratings diverge for tokens paired with African American faces being far more inaccurate. Those with a more neutral attitude rated tokens paired with African American faces as more accurate and those with a more pro-white attitude rated the same tokens paired with the same African American faces as more inaccurate. In the other two

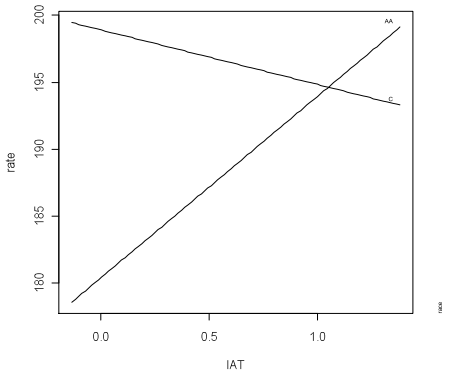
conditions, there was a generic tendency for people with higher IAT scores to rate both tokens paired with African American and Caucasian faces as less accurate. In summary, the interaction between visual biasing and implicit attitudes was most robust for the condition in which stimuli were gated at the burst.

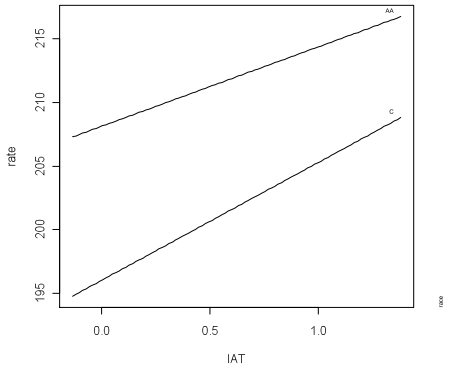
Figure 4.11. Interaction between accuracy ratings of speech tokens when paired with a photograph of an African American (line labeled AA) girl, African American boy, Caucasian (line labeled C) girl, or Caucasian boy and Implicit Association Task score (IAT) as determined by a LME analysis of Stop Type data. (Girls' data is on the left, boys' data is on the right. Lower accuracy rating values, along the y-axis, indicate listener judged production as more accurate. Lower IAT scores, along the x-axis, indicate neutral racial attitude; higher IAT scores indicate pro-white bias.)

Tokens paired with Photos of Girls

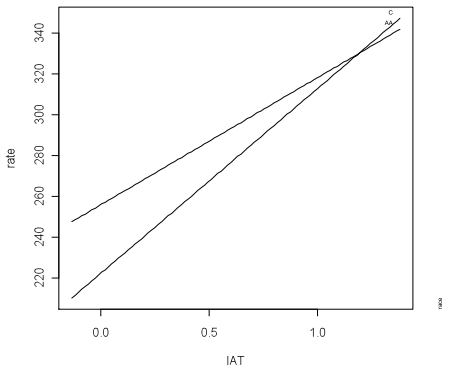
Tokens paired with Photos of Boys

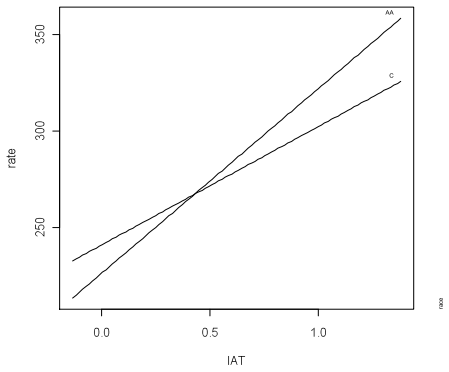
Word Final /t/ Unaltered



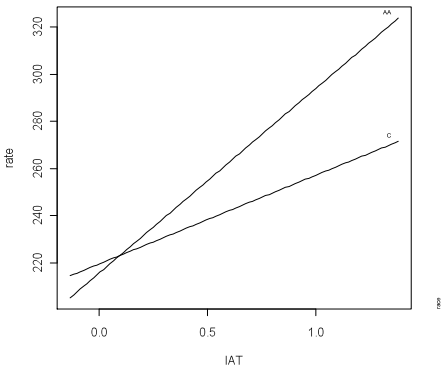


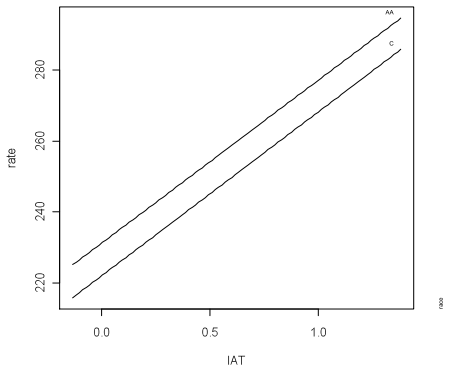
Word Final /t/ Gated at Burst





Word Final /t/ Gated 20 ms before Burst

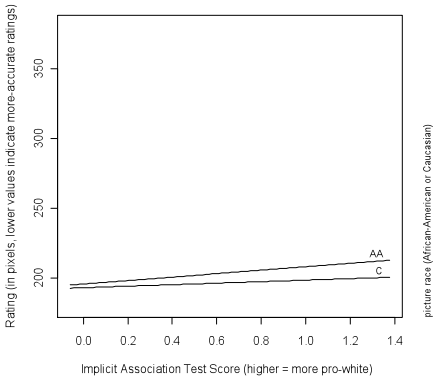
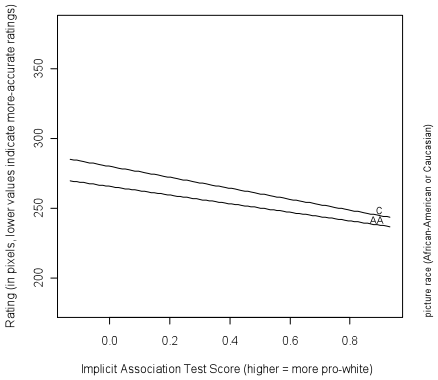
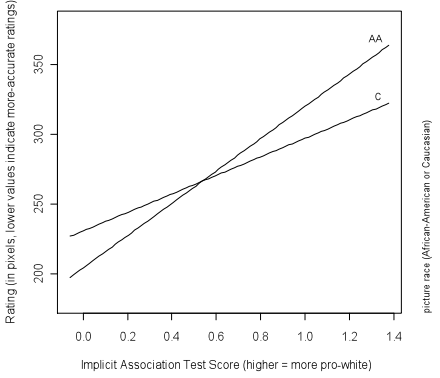
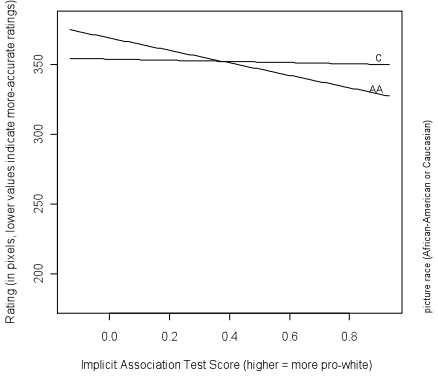
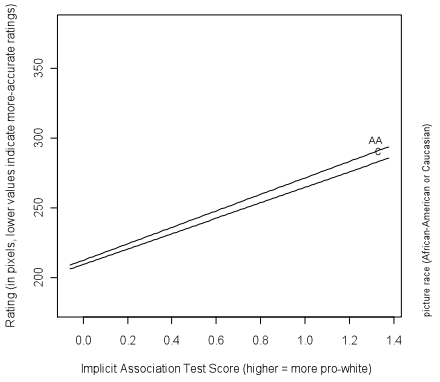
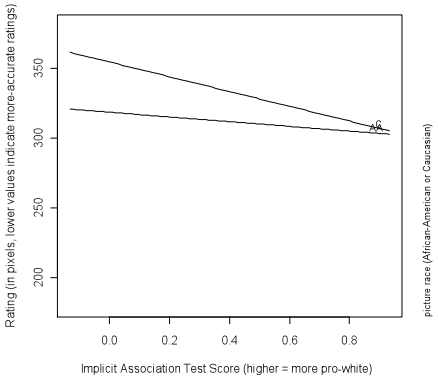




The last two interactions were four-way interactions between IAT scores, by Photograph Race, by Child Time, by Stop Type-Whole [β=31.165, t(3925)=1.952, p=0.0490] and IAT scores, by Photograph Race, by Child Time, by Stop Type-20 ms [β=27.861, t(3925)=1.911, p=0.0524]. Illustrated in Figure 4.12, the graphs are separated by experience with children and stop condition. The inexperienced listeners (those who 0 to 3 years experience on a self-report survey question asking "How much time per week do you spend with children?" with 0 = none, and 10 = extremely frequently) are on the left, and the experienced listeners (those who rated 4 years or more) are on the right. The first set of graphs represents the whole word condition, the second set portrays the closure condition, and the final set depicts the minus 20 ms condition. Looking at the experienced listeners: the relationship between IAT scores and accuracy ratings grouped by the race of the photograph the tokens were paired with was generally what we expected to see: the greater the pro-white bias, the more likely listeners were to rate tokens paired with photographs of African American children as less accurate than tokens paired with photographs of Caucasian children, except for the most altered condition where the stop was gated 20 ms before the burst where ratings essentially converged. The experienced listeners present a more varied interaction. It is important to note that more experienced listeners exhibited a smaller range IAT scores overall implying that overall, these participants' implicit attitudes about race (black/white) are more neutral. They also demonstrate a completely opposite pattern of IAT-by-accuracy ratings than did the inexperienced listeners: those with higher IAT scores were more likely overall to rate talkers as more accurate rather than less accurate. Finally, in the two altered conditions

(closure and minus 20 ms), more experienced participants displayed an opposite than expected pattern: listeners with a greater pro-white bias actually rated tokens paired with African American faces as more accurate than tokens paired with Caucasian faces. In summary that the interaction between visual cueing and implicit attitudes was most robust for the low-experience listeners.

Figure 4.12. Interaction between accuracy ratings of speech tokens with word-final /t/ unaltered, gated at the burst, and gated 20 ms before the burst when paired with a photograph of an African American (line labeled AA) or Caucasian (line labeled C) child and IAT divided by experience with children in years (Child Time) as determined by a LME analysis of Stop Type data. (Data separated by experience with children. Less experienced is defined as 3 or less on a self-report scale of 1 to 10-1 being 'No Experience' and More experienced is defined as 4 or more on a self-report scale of 1 to 10-10 being 'Extremely Frequent Experience'. Lower accuracy rating values, along the y-axis, indicate listener judged production as more accurate. Lower IAT scores, along the x-axis, indicate neutral racial attitude; higher IAT scores indicate pro-white bias.)

	Less Experience with Children (Self-report rating of ≤ 3)	More Experience with Children (Self-report rating of ≥ 4)
Word Final /t/ Unaltered		
Word Final /t/ Gated at Burst		
Word Final /t/ Gated 20 ms before Burst		

Discussion

We designed this study to look at how the perception of race affects the perception of children's speech. We were interested specifically on the influence that the perceived race of a talker could have on how adults judge the accuracy of children's speech. We were further interested in the impact of implicit social attitudes and experience on the perception of child's speech accuracy, and whether these factors interact. The research on the role of social knowledge in speech perception is a relatively new area of inquiry and to date the majority of the research has looked at adult speech. This study extended this research to the perception of children's speech. This extension is critical in light of the fact that perceptual judgments of children's speech are regularly used by speech-language pathologists to determine whether children have normal or atypical articulation abilities.

First, we predicted that the listeners would rate the accuracy of speech differently when it was paired with photographs of African American children's faces than when it was paired with photographs of Caucasian children. Our two analyses resulted in slightly different findings regarding this prediction. The results of an ANOVA showed, contrary to expectations, that race significantly affected judgments of the accuracy of words containing initial fricatives but not judgments of words with final /t/. However, a set of linear mixed-effects models showed effects in the predicted direction: words were rated as less accurate when they were paired with pictures of African American children than when paired with pictures of Caucasian children. This discrepancy across analyses was likely due to the fact that the LME allowed us to examine whether experience perceiving

children's speech, and implicit attitudes about race, affected ratings. The LME showed that ratings of words with final stop consonants were indeed mediated by both of these factors. The mediation was such that the individuals with more experience perceiving children's speech rated words paired with African American children's faces as more accurate than those paired with Caucasian children's faces. Individuals whose performance on an Implicit Association Test suggested more pro-white attitudes rated words with final /t/ to be more accurate when they were paired with Caucasian children's faces than when they were paired with African American children's faces. Individuals whose IAT scores showed a more race-neutral attitude showed the opposite effect. As described in the results section, there was a very complex interaction between this tendency, the gender of the child whose picture was paired with the speech, the amount of experience the listener had perceiving children's speech, and whether the token was unaltered or contained a simulated error. The general conclusion we can draw from what we observed is that the interaction between IAT score and race-biasing in accuracy ratings is strongest for /t/-final tokens that were gated at the burst, when paired with pictures of girls' faces, and when rated by listeners with relatively little experience perceiving children's speech. The /t/-final tokens gated at the burst imitated an unreleased final /t/. This condition emulates a dropped final consonant which would likely be rated as an error or by a speaker of mainstream American English, but a correct production by a speaker of African-American English. That is, ratings by inexperienced listeners with relatively more pro-white implicit attitudes were biased most when

listening to a phonological form whose accuracy was ambiguous, and which they thought to be produced by a girl.

We predicted that, because of their training, the graduate students would have explicit knowledge of deleted final consonants as a feature African American English dialect. Therefore, they would rate tokens with altered final stops paired with photos of African American children as more accurate than when paired with photos of Caucasian children. We predicted that undergraduates would show no such discrepancy in their ratings. The results of the ANOVAs showed some significant effects of listener group. In general, the listener groups were more likely to be affected by the acoustic manipulations to give the illusion of different types of errors. However, the ANOVAs did not show an interaction between listener group and the race of the face with which tokens were paired as predicted.

A few other findings bear mention. First, the tokens altered to imitate a speech error were rated to sound significantly more inaccurate than the speech tokens altered to imitate a correct production, illustrating that our manipulation convincingly emulated common speech errors. Next, the IAT did predict overall accuracy ratings for the fricative tokens, in that accuracy ratings made by participants with a greater pro-white bias were more extreme than those made by participants with a more neutral attitude about race. While not lending itself to an obvious theoretical interpretation this finding is noteworthy in that individuals with higher IAT scores had more polarized ratings of words with /s/ and /θ/ than did those with lower IAT scores. A similar interaction was not found between self-reported experience perceiving children's speech and ratings,

suggesting that this link is not a consequence of the amount of exposure to children's speech, but instead reflects some variable that itself is correlated with the IAT, which remains to be uncovered.

Why did those listeners with higher IAT scores (pro-white bias) rate tokens paired with an African American child's photo as more inaccurate than tokens paired with a Caucasian child's photo? One interpretation is that this reflects unconscious racial stereotypes held by the listener. It is important to note that the great majority of our participants self-identified as Caucasian, thus, this result may also be due to unconscious in-group/out-group biases. Participants with more neutral attitudes toward race did not rate tokens paired with Caucasian faces as more accurate as predicted. In fact, they rated tokens paired with African American faces as more accurate than those paired with Caucasian faces.

The outcome increased in complexity when gender was considered alongside IAT score and race. Tokens paired with photos of boys followed the predicted pattern. Listeners with pro-white biases rated tokens paired with pictures of African American boys as less accurate than when paired with pictures of Caucasian boys and those with more neutral attitudes about race rated those same tokens as more accurate when paired with pictures of African American boys. However, pairing the tokens with girls' photos reversed the ratings for both groups, pro-white and neutral. Since boys produced the original tokens, it is possible that some of the tokens were not gender neutral and therefore confounded the results. Respondents may have rated the tokens not on accuracy of phonemic production but rather gender typicality. This remains an unknown

component of our study, as we did not measure implicit attitudes about gender and therefore gender stereotypes may have played a role. Furthermore, treating gender and race as independent fixed factors may be problematic. For instance, stereotypes of a white female and a black female may have little overlap. These factors would need to be considered in future research. Further refinement of an instrument such as the IAT may be necessary to take into account more complex identity configurations when measuring implicit biases. As an example of the complexity of interactions between implicit attitudes about gender and race, when we added stop type to the analysis, gender appeared to have a powerful interaction with the other fixed effects. However, we could not discern a meaningful pattern that lent itself to theoretical explanation.

Experience with children did appear to conform to our predictions. Graduate students in our study reported an average of 4.2 years of experience with children as opposed to undergraduates averaging 2.7 years of experience with children. Those participants who reported having more experience with children rated the tokens with final stops as more accurate when paired with a photo of an African American child while those with less experience with children rated those tokens as less accurate when paired with an African American face. Experience with children may have afforded these participants with implicit knowledge of African American English (AAE) dialect and resulted in the higher accuracy ratings when tokens were paired with African American faces. However, we did not gather data to confirm this possibility. Future studies would do well to implement ethnographic research in order to gather a more nuanced and detailed picture of type and breadth of experience. Importantly, this interaction did not

occur with tokens with altered fricatives. This may support our hypothesis in that substitution of /θ/ for /s/ is not a feature of AAE.

We observed several significant effects and interactions that serve as evidence to support previous research on the impact of implicit social attitudes and knowledge on speech perception. While effect sizes were small, this does not necessarily diminish their importance. We do not know the impact even small effects may have on the judgments of speech production accuracy in the field.

While our study found evidence to support many of our conclusions, there are several aspects of the study that could be changed to further elucidate the patterns we observed. Our speech stimuli were all produced by Canadian boys, some more gender neutral than others. Ideally, we would compile a more representative corpus of speech tokens from both girls and boys of many ages, from different dialectical regions, and of both Caucasian and African American groups. Future studies should include participants with a wider range of experience with children as well as more detailed measures of experience. We knew no details of the experiences our participants had with children. For instance, our subject pool should include participants who have experience with children from a wide variety of cultural and socio-economic backgrounds.

Concerns may be raised about the ecological validity of the subjects in our study. It could be argued that our participant pool was a select and non-random sampling of the general population. Thus, the data collected would not necessarily represent the population at large. However, we were interested in the impact of biases on perception of children's speech in the Speech Language clinical setting. The subjects in our study were

of similar racial profile, socio-economic status and educational level to the typical Speech Language Pathologist. However, children interact with adults from much more varied backgrounds. For instance, day care workers who spend a great deal of time with children during a most critical stage of language development are not well represented in our study. We would like to expand our participant pool to include representatives from all groups who work closely with children.

Another issue with representation in the participants in our study occurred with the IAT. All participants fell within a neutral attitude to pro-white bias spectrum in IAT score. No participant scored in the pro-black range. Those participants who scored in the pro-white bias range also tended to rate all speech tokens as more inaccurate than those who fell in the neutral category. It is possible that people who score in the more biased ranges of the scale are simply more extreme in their use of scaled instruments. They may tend to gravitate towards the ends of a scale when recording judgments. We cannot know if this is the case without a group of participants who fall within the pro-black range.

The IAT has been a useful instrument to measure implicit bias. However, there are severe limitations to treating macro-sociological characteristics like race as a fixed and independent category. Characteristics such as race, class, gender, and heritage interact in complex ways as evidenced by our results. New measures that account for the multivalent ways in which these categories relate with and mediate each other must be created if we are to better understand these complex interactions. The IAT is limited by its dichotomous nature, i.e. good/bad, black/white. A tool that accounts for multiple categories such as black/yellow/brown/white would be more ecologically valid.

We employed a relatively new statistical model: the linear mixed effects model. An advantage of this model is that it does not oversimplify nuanced interactions between complex effects. However, it does not offer a measure of effect size when effects and interactions occur, limiting our ability to interpret the implications of findings. Continued improvement and utilization of more precise and nuanced statistical models is particularly essential to this area of inquiry.

Finally, much of the research that has been conducted in this area of the effects of indexical knowledge on speech perception has been limited in its scope. Continued expansion into areas beyond sociophonetics is critical at this juncture in the research conversation.

Evidence is mounting that implicit biases can affect perception of speech; however we do not know the full impact of attitude and experience in the assessment of children's speech. And a larger question remains unanswered: what is our response to these effects in real world settings? Can these effects be mitigated? What procedures might we implement in the clinical setting?

Is an implication of this line of inquiry that we need to subject people who work with children to tests that measure implicit biases with the intention of mitigating any negative consequences that may result from consequences of these biases may inflict? Do we commit resources to developing programs to decrease the impact of social bias on language acquisition, assessment, and learning? These are questions that remain to be considered by the community at large.

Bibliography

Baayen, R.H., D.J. Davidson, & D.M. Bates (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.

Babel, M.E. (2009) *Phonetic and Social Selectivity in Speech Accommodation. Dissertation*. University of California, Berkeley.

Boersma, P, Weenik D. (2005). Praat: Doing phonetics by computer (Version 4.3. 27)[Computer program]

Campbell-Kibler, K. (2006). *Listener Perceptions of Sociolinguistic Variables the Case of (Ing)*. Dissertation. Stanford University.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd Edition). New York: Routledge Academic.

Crocker, L.P. (2006). *Speech and Voice Characteristics of Boys with Gender Identity Disorder. Unpublished B.A. Honors Thesis.* Department of Speech-Language-Hearing Sciences, University of Minnesota, Twin Cities.

Crocker, L.P., & Munson, B. 2006. Speech characteristics of gender-nonconforming boys. Presentation given at the Conference on New Ways of Analyzing Variation in Language, Columbus, OH.

Dasgupta, N. & Greenwald, A.G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, 81(5), 800-814.

Dasgupta, N., McGhee, D.E., Greenwald, A.G., Banaji, M.R. (2000). Automatic Preference for White Americans: Eliminating the Familiarity Explanation. *Journal of Experimental Social Psychology* 36, 316-328.

Drager, K. (2005). From Bad to Bed: The Relationship Between Perceived Age and Vowel Perception in New Zealand English. *Te Reo* 48, 55-68.

Eyewire Images (2002). *Photography: Babies*. Getty Images, Inc.

Greenwald, A.G., McGhee, D.E., & Schwartz, J. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74:1464–1480.

Greenwald, A.G., Nosek, B., & Banaji, M.R. (2003). Understanding and Using the Implicit Association Test: I. An Improved Algorithm. *Journal of Personality and Social Psychology*, 85:197-216

Hay, J. and K. Drager. (2007). Sociophonetics. *Annual Review of Anthropology*, 36, 89-103.

Hay, J., P. Warren, & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics,* 34, 458-484.

Jelenec, P. & Steffens, M.C. (2002). Implicit Attitudes Toward Eldery Women and Men. *Current Research in Social Psychology*, 7(16), 275-292.

Labov, William. (1972). *Language in the Inner City: Studies in the Black English Vernacular.* Philadelphia: University of Pennsylvania Press.

Lindblom, B. (1990). Explaining variation: A sketch of the H and H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling.* Kluwer.

McGurk, H. & MacDonald J. (1976). Hearing lips and seeing voices. *Nature,* 264, 746-748.

Meyerhoff, M & Niedzielski, N (2000). The globalisation of vernacular variation. Journal of Sociolinguistics 7(4), 534 - 555

Munson, B. & Seppanen V.R. (2009). Perceived Suggested Gender Affects Ratings of the Quality of Children's Spoken Narratives. Presentation. http:www/tc.umn.edu/~munso005

Munson, B., Edwards, J., Schellinger, S.K., Beckman, M.E., & Meyer, M.K. (2010) Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of *Vox Humana*. *Clinical Linguistics and Phonetics*, 24(4-5), 245-260.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, 18, 62-85.

Nygaard, L.C. & Pisoni, D.B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics* 60(3), 355-376.

Nygaard, L.C. & Lunders, E.R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory and Cognition* 30(4), 583-593.

Palmeri, T.J., Goldinger, S.D., and Pisoni, D.B. (1993). Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words. *Journal Experimental Psychology: Learning Memory, and Cognition*. 19(2), 309-328.

Perkall, J.S. & Klatt, D.H. (Eds.) (1986). *Invariance and Variability in Speech Processes* Psychology Press

Rubin, D. L. (1992). Nonlanguage Factors Affecting Undergradutes' Judgments of Nonnative English-speaking Teaching Assistants. *Research in Higher Education* 33(4), 511-531.

Schellinger, S.K., Edwards, J. & Munson, B. (2010). The Role of Intermediate Productions and Listener Expectations on the Perception of Children's Speech. Unpublished Manuscript.

Schneider, W., Eschman, A., Zuccolotto, A. (2002) *E-Prime User's Guide, version 1.2 [computer program]* Pittsburgh: Psychology Software Tools.

Smit, A.B., Hand, L., Freilinger, J.J., Bernthal, J.E., and Bird, A. (1990). The Iowa Articulation Norms Project and its Nebraska Replication. *Journal of Speech and Hearing Disorders,* 55, 779-798.

Staum Casasanto, L. (2008). Does Social Information Influence Sentence Processing? *Annual Meeting of the Cognitive Science Society*. 799-804

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., & Sedivy, J.C. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science*, 268, 1632-1634.

Thomas, E.R. (2007). Phonological and Phonetic Characteristics of African American Vernacular English. *Language and Linguistics Compass,* 1(5), 450-475.

Walker, S., Bruce, V., O'Malley, C. (1995). Facial identity and facial speech processing: familiar faces and voices in the McGurk Effect. *Perception and Psychophysics,* 57(8): 1124-1133.

Appendix A: Debriefing

Now that you have finished this experiment, we would like to tell you what we are studying. In this experiment, we are interested in whether people perceive speech differently depending on the race of the person who produced it. In the task in which you were rating children's speech, you sometimes viewed pictures of African American children, and sometimes children who were Caucasian or Asian-American. We are interested in whether people rate speech differently in these different conditions. We are also interested in whether this tendency is related to people's experience perceiving speech, which is why we are comparing untrained people's perception to the perception of trained speech-language pathologists. We are also interested in how this interacts with people's unconscious perception of race, which is what we measured in the reaction-time task.

Our goal is not to label people or to judge their behavior. Instead, we are interested in finding the best way to assess children's speech production so that we can make sure that what people report is as close as possible to what the children do.

Now that we have told you what our purpose is, we would like to give you the opportunity to withdraw your data. You can do this without changing your compensation, and without affecting your relationship to the University or to the people who did this study. If you choose to leave your data in, please know that we will treat it completely confidentially. We will not link your name to your experimental results in any published report of this work. All of our analyses will be done using generic subject IDs. All of the records linking your name to the subject ID are kept in a locked cabinet in

a locked research lab, and will be accessible only to the researchers who are working on this project. If you choose to leave your data in now, but change your mind later, you may still remove your data. Just contact us at Benjamin Munson at Munso005@umn.edu, or (612) 624-0304.