LSHSS

Research Article

Does Speaker Race Affect the Assessment of Children's Speech Accuracy? A Comparison of Speech-Language Pathologists and Clinically Untrained Listeners

Karen E. Evans,[a] Benjamin Munson,[a] and Jan Edwards[b]

Purpose: Some pronunciation patterns that are normal in 1 dialect might represent an error in another dialect (i.e., [koʊl] for *cold,* which is typical in African American English [AAE] but an error in many other dialects of English). This study examined whether trained speech-language pathologists and untrained listeners accommodate for presumed speaker dialect when rating children's productions of words. This study also explored whether effects of presumed race on perceived speech accuracy are mediated by individuals' knowledge and beliefs about AAE and their implicit attitudes about race.
Method: Multiple groups of listeners rated the accuracy of a set of children's productions of words that have a distinct pronunciation in AAE. These were presented in 1 of 3 conditions: paired with no visual stimulus (to assess baseline accuracy) or paired with either African American children's faces (to suggest that the speaker uses AAE) or European American children's faces (to suggest that the speaker does not use AAE). Listeners also completed a set of measures of knowledge and attitudes about AAE and race, taken from previous studies.
Results: Individuals in both groups rated children's productions more accurately when they were presented with African American children's faces than when paired with European American faces. The magnitude of this effect was generally similar across the 2 groups and was generally strongest for words that had been judged in the baseline condition to contain an error. None of the individual-differences measures predicted ratings.
Conclusions: Assumptions about speaker attributes affect individuals' assessment of children's production accuracy. These effects are robust across trained and untrained listeners and cannot be predicted by existing measures of knowledge and attitudes about AAE and race.

As the cultural and linguistic diversity of the United States continues to increase, no population reflects these demographic shifts so dramatically as that of young children. For the first time, more than half of children under 1 year of age in the United States are non-White, Latino/a, or both (U.S. Census Bureau, 2012). Many of these children arrive at school speaking languages other than English, whereas others are speakers of nonmainstream English dialects, such as African American English (AAE). To assess the speech and language skills of all children appropriately, speech-language pathologists (SLPs) must understand the effect their own biases may have on their clinical practices, including the very basic practice of perceiving and denoting productions of sounds and words. In its official statement on the "Knowledge and Skills Needed by Speech-Language Pathologists and Audiologists to Provide Culturally and Linguistically Appropriate Services" (American Speech-Language-Hearing Association [ASHA], 2004), ASHA lists knowledge of the "influence of one's own beliefs and biases in providing effective services" first among many requisites (ASHA, 2004: 1.1). However, the extent of this influence is not yet well understood, especially when one considers the particular influence that listener bias may exert on speech perception.

Skilled speech perception is a core competency for SLPs serving clients with phonological and articulation disorders. Norm-referenced articulation assessments are scored according to impressionistic judgments made by the examiner. SLPs use these assessments regularly to screen

[a]Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis
[b]Department of Communication Sciences and Disorders, University of Wisconsin–Madison
Correspondence to Benjamin Munson: Munso005@umn.edu

Editor-in-Chief: Shelley Gray
Editor: Ignatius Nip
Received October 29, 2017
Revision received February 7, 2018
Accepted February 27, 2018
https://doi.org/10.1044/2018_LSHSS-17-0120

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

for and diagnose speech sound disorders in children and to determine eligibility for clinical services. If conscious or unconscious listener biases affect even a fraction of test items, the results of the test could change. This in turn could have consequences for determining whether a child has an impairment and is eligible for clinical services and for the selection of specific sounds or contrasts to treat.

Speech perception is a highly complex skill, characterized by the fundamental lack of invariance between the acoustic forms of speech and the perceptual labels that listeners apply (Perkall & Klatt, 1986). Part of the complexity of speech perception comes from the many sources of information that individuals rely on when making an association between an acoustic signal and the message being communicated (Kleinschmidt & Jaeger, 2015). This includes "bottom-up" information about the acoustic structure of the message, such as the association of an interval of high-frequency aperiodic noise with the phoneme /s/. There is also "top-down" information, such as the knowledge that the peak frequency for /s/ is lower when it precedes a sound made with lip rounding than when it does not. This top-down knowledge helps listeners calibrate their expectations, so that an interval of frication that might be labeled as /s/ in a rounding context might be labeled as the phonetically similar sound /ʃ/ in a nonrounding context. This example is but one of many. Research has documented that there are myriad ways that listeners overcome the invariance problem to disambiguate acoustic information and establish a steady percept. These include implicit knowledge of coarticulation (Lindblom, 1990), the effects of speaking rate (e.g., Miller, 1981), and emotional tone of voice (Nygaard & Lunders, 2002).

One major advance in speech perception research in the past 20 years is the finding that listeners' expectations about socially meaningful linguistic variation constrain speech perception. One illustration of this is the finding that a sound intermediate between /s/ and /ʃ/ can be perceived as either more /s/-like or more /ʃ/-like depending on whether the talker believes it to have been produced by a man or by a woman (Strand & Johnson, 1996). This finding has been replicated for a variety of social variables and a variety of speech sounds. These findings invite a systematic investigation of whether inferences about speakers influence the way that children's speech is perceived. The broader goal is to understand whether such influences, if any, affect the assessment of children's speech clinically. This work is consistent with decades of research on the effect of visual information on speech perception. The classic McGurk effect (McGurk & MacDonald, 1976) demonstrates that the articulatory gestures a listener sees will alter what phoneme he or she perceives, even given prior knowledge of the stimulus (Walker, Bruce, & O'Malley, 1995).

Characteristics of the talker also interact with perception in both bottom-up and top-down directions. In the former case, listeners often form assumptions about an individual on the basis of speech patterns, such as when one infers a person's gender and approximate age over the phone. In the latter situation, known or assumed information about the talker leads to shifts or corrections in perception. This is illustrated by an adult's ability to understand young children with developmental phonological errors (e.g., substitution of /w/ for /r/ as in /wæb t/ for *rabbit*). A growing body of research has examined these relationships with respect to various indexical characteristics of the talker, including age, gender, race, and region of origin, among others. For example, Munson, Edwards, Schellinger, Beckman, and Meyer (2010) found effects of presumed age upon perception in investigations of child speech. In that study, adult listeners rated the accuracy of productions of /s/ ranging from correct /s/ to misarticulated /θ/. When the listeners believed that the children were older, they rated productions of /θ/ as more accurate than if they believed the speakers were younger.

Speech perception also intersects with stereotypes related to social and cultural variation in various ways. Niedzielski (1999) asked residents of Detroit to match from a selection the vowels they perceived in the speech of another Detroit resident, whom they were led to believe was either from Detroit or from Canada. The speaker's dialect, typical of middle-class residents of Detroit, included the raised-diphthong vowels seen in the Northern Cities Chain Shift dialect. Listeners who believed that the speaker was Canadian correctly identified the raised-diphthong vowels, but listeners who believed that the speaker was a fellow Detroiter perceived the diphthongs as falling closer to mainstream American English (MAE) unraised forms. These findings support the notion that social information invokes expectations that, in turn, affect speech perception. Furthermore, the fact that Detroit residents perceived the speech of a member of their community as closer to standard than that of an outsider suggests that attitudes about nonstandard dialects also affect perception. Similar findings on cross-dialect speech perception are presented by Hay, Nolan, and Drager (2006), who examined how speaker age, sex, and socioeconomic status (SES) affect the perception of vowels in New Zealand English.

Knowledge or assumptions about a speaker's race or ethnicity and related listener attitudes about this also come to bear on the listener's perceptions. Participants in Rubin (1992) rated the speech of the same native speaker of American English as sounding more "foreign" or nonstandard when it was paired with an Asian face than when it was paired with a European American face. More recently, Staum Casasanto (2008) designed a response time task to investigate how listeners use presumed race and knowledge of social dialects to disambiguate sentences. Staum Casasanto focused on the perception of AAE, a dialect of English observed throughout the United States and one that historically has been subject to significant controversy (e.g., Lakoff, 2000). AAE is a distinct dialect of English characterized by phonological, morphological, and syntactic differences from "standard" or "mainstream" American English, although it shares many features with varieties of English spoken by European Americans in the southern United States (see, e.g., Rickford & Rickford, 2000, for a review). Although African Americans have historically been the chief group of AAE speakers, they are not the only speakers of this

dialect; conversely, not all African Americans speak AAE. This leads to a complex relationship between individuals' attitudes toward race and their attitudes toward AAE. According to the U.S. Department of Education (2001), African American students are consistently overrepresented on special education caseloads, so the issue of adults' perception of AAE in children is critical.

Although there has not yet been a large-scale study of AAE using population-based sampling, results from smaller scale studies have led to catalogs of these distinctive features. Thomas (2007) provides a review of phonological features of AAE. These include the glottalization of final /t/, the reduction of final consonant clusters, and the vocalization or deletion of final /l/. AAE also has many distinctive morphosyntactic features, such as the nonuse of /s/ to mark regular third-person singular morphology in verbs or the possessive form of nouns. Some of these features are very salient to speakers of English, even being encoded in nonstandard spellings of words or folk-linguistic descriptions of AAE (i.e., the spelling of *sho'* for *sure,* presumably reflecting the AAE nonrhotic pronunciation [ʃo]).

Previous studies have provided mixed findings regarding the extent to which listeners can identify whether a speaker is African American from phonetic cues alone. Thomas and Reaser (2004) reviewed studies on this topic. They report some studies that showed good identification of race when relatively long stretches of speech are used. One recent study illustrating this is presented by Gaither, Cohen-Goldberg, Gidney, and Maddox (2015), who found that biracial speakers' race was perceived differently in audio-only samples collected in two conditions, one in which a White identity was primed and one in which a Black identity was primed. These ratings were made from 10- to 20-s stretches of speech whose content was not expected to cue racial identity. In contrast, the evidence that listeners can perceive race from single-word productions is weak. Lass, Tecca, Mancuso, and Black (1979) found that speakers' ethnicity was guessed accurately only 55% of the time when single-word stimuli were used.

Staum Casasanto (2008) studied the effect of race on the perception of final /st/ clusters. Where speaker of MAE dialects fully produce final /st/ clusters, speakers of AAE might produce a singleton /s/. This may lead to the neutralization of lexical contrasts, such that *mass* would be indistinguishable from *mast* when produced by a speaker of AAE but not by one of MAE. Staum Casasanto presented listeners with words embedded in a phrase, such as "The [mæs] probably lasted…" in which the stimulus could be either "mass" or "mast" depending on the following phase. The carrier phrases were paired with pictures of either African Americans or European Americans. Afterward, listeners read a phrase that completed the sentence, and response time was measured as they judged whether or not the sentence made sense. Some completed sentences made sense if the stimulus word had a reduced final cluster, as in "The [mæs] probably lasted through the storm," whereas others made sense only without a reduced cluster, as in "The [mæs] probably lasted an hour on Sunday." Response times differed significantly depending on the race of the picture listeners saw and the word form that correctly completes the sentence. If listeners saw a European American face, they were quicker in responding to sentences with no necessary cluster reduction, whereas if they saw an African American face, they were faster in their responses to sentences where cluster reduction would be needed. This result also highlights the conflation of race with dialect in the minds of many listeners; seeing an African American face led listeners to associations with AAE, although not all African Americans speak that variety of English.

The current study is inspired by Staum Casasanto's work. It examines effects of speaker race on speech perception both by laypeople and by people with clinical training in speech-language pathology. The goal of this study was to investigate adults' judgments of the speech of children who speak AAE, including the effects of presumed speaker race, the differences between judgments of speech-language clinicians and lay listeners, and the influence of attitudes and beliefs on these effects. An audiovisual (AV) perception experiment was conducted in which clinicians and untrained listeners rated the accuracy of words spoken by children they believed to be either African American or European American. For the current study, the stimuli were speech samples from a large number of young AAE-speaking children of both sexes. These samples varied widely in articulatory accuracy and the presence of features that are characteristic of AAE production by children. They included four contrasts that are characteristic of AAE: vocalization of final /l/, omission of /d/ from final /ld/ clusters, omission of /s/ from plural forms, and glottalization or deletion of final /t/. On the basis of previous research, we predicted that, as a group, listeners would be more likely to rate productions with AAE-typical pronunciations as correct when they were paired with an African American child's face than when paired with a European American child's face. That is, we predicted that listeners would have implicit knowledge of AAE-related variation and that this would lead them to rate speech differently depending on the presumed race of the speaker. We also predicted that there would be considerable individual differences in the extent to which individual ratings differed across listeners. We predicted that there would be a bigger effect of presumed race on ratings by trained SLP, as these individuals have coursework that teaches the features of AAE. We also predicted that there would be a bigger effect in individuals with better knowledge of AAE, as assessed objectively, and more positive attitudes toward African Americans, measured through the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998) and through responses on a questionnaire.

Method

Overview

The study consisted of three parts: two speech-rating tasks, an implicit association task designed to measure implicit racial bias, and a multipart questionnaire designed to

assess knowledge and attitude toward nonmainstream dialects. In the first speech-rating task, listeners rated the accuracy of speech samples from children who speak AAE without any information about their race. This task (baseline rating task) was used to determine the baseline-perceived accuracy of each token without suggesting anything about the speaker's race. It was used as a finer grained and more ecologically valid measure of accuracy than is given by phonetic transcriptions, as it was based on the average judgments of a panel of listeners using a continuous rating scale (see Schellinger, Munson, & Edwards, 2017, for a discussion of the utility of panel ratings), rather than the transcriptions of a small number of highly trained listeners using a categorical rating system. Some of the phonetic variants examined in this study, such as the difference between a fully articulated final /l/ and a fully vocalized one, are intrinsically continuous, and are thus not captured well by phonetic transcriptions using a discrete symbol set. Moreover, there is ample phonetic evidence that there are developmentally meaningful phonetic differences among sounds that have been transcribed to be accurate (e.g., Holliday, Reidy, Beckman, & Edwards, 2015; Romeo, Hazan, & Pettinato, 2013). The perceptual ratings in the baseline task were intended to ameliorate that weakness.

In the second speech-rating task (AV rating task), a different group of participants rated these same tokens, paired with pictures of children who were either African American or European American and who the listeners were told had produced the speech they were hearing. This task was designed to assess the influence of perceived race on ratings of the accuracy of children's speech. In both tasks, participants rated the accuracy of the speech on a continuous visual analog scale (VAS). The listeners in the AV rating task also completed the IAT (Greenwald, McGhee, & Schwartz, 1998), previously created for investigations of racial attitudes. The AV rating task listeners also completed the multipart questionnaire, which was designed to elicit information pertaining to participants' knowledge of the features of AAE, their explicit attitudes about AAE and other issues related to dialects of English, and, for the trained listener group, information about their experience working with AAE-speaking clients. We examined whether the IAT and the questionnaire measures predicted the influence of race on ratings of children's speech accuracy.

Participants

A total of 60 adults participated in this study. Twenty adults (all of whom were untrained, i.e., they had no formal education in speech-language pathology) completed the baseline speech-rating task. Forty adults (20 trained SLPs, 20 untrained listeners) participated in the AV speech-rating task. Participation was restricted to native monolingual English speakers over the age of 18 years with no history of speech, language, or hearing disorders (other than articulation errors affecting only a small number of sounds), on the basis of self-report. Second-language proficiency attained after childhood was not a disqualifying factor. Recruitment materials described the experiment as a "speech perception study" and included no mention of race or AAE. Participants were compensated $10 for their participation.

Twelve practicing pediatric SLPs (11 female, one male) and eight advanced graduate students (all female, in their last semester of study in speech-language pathology) from the Minneapolis–St. Paul area formed the trained listener group for the AV rating task. Seventeen out of 20 trained listeners self-identified as European American, one as Asian American, and two as "other." The mean age of the trained listeners was 31.79 years (SD = 10.90), and the group reported a mean of 9.1 years of experience as SLPs (SD = 11.32). All of the trained listeners were either currently enrolled in or had graduated from a graduate program in speech-language pathology accredited by ASHA's Council on Academic Accreditation. Council on Academic Accreditation–accredited programs require students to document knowledge and skills related to speech-language pathology, including knowledge of dialect variation in speech. The advanced graduate students had taken a course with the second author in which AAE variation was discussed explicitly.

Two groups of 20 individuals were recruited from the University of Minnesota community to form the untrained listener groups for the baseline rating task and the AV rating task. Data were excluded from one additional participant in the AV rating task, who, after completing the study, reported that she had consistently reversed her answers during the speech-rating task. Thirty-five out of 40 untrained listeners self-identified as European American, one as African American, two as Asian American, and two as "other." The mean age of the untrained listeners was 22.60 years (SD = 5.83). The two listener groups assigned to the baseline condition and those assigned to the AV condition did not differ in age or in gender composition.

Speech-Rating Task

Stimuli

The speech samples for the speech-rating tasks were single-word productions that were collected at the University of Wisconsin–Madison as part of a study of children's perception of dialect variation. The results of that study can be found in Edwards et al. (2014). Speech samples were elicited from 109 African American children between the ages of 4 and 9 years. Most children in the sample were from low-SES households (on the basis of maternal education level and total family income). The presence of dialect features in the children's speech was identified through a 50-utterance language sample. The children used at least one AAE feature in those language samples, as described in Edwards et al. (2014), which also describes these children's performance on standardized language measures.

The single-word samples used here were separate from those samples. They were single-word productions collected as part of the familiarization phase of a word comprehension

experiment. The stimuli consisted of eight practice items and 18 pairs of pictureable words that would be familiar to the children: nine singular/plural pairs, such as *hat*/*hats,* and nine monomorphemic word pairs differing in the presence of a final consonant cluster, such as *goal*/*gold*. A female speaker of AAE recorded each word embedded in the carrier phrase "*Say [word], please*." These recordings were paired with color photographs and presented to the children on a touch screen. Children's productions were audio-recorded.

From the 44 monosyllabic words elicited from the speakers, words with four different final consonant patterns were selected: two singletons (final /t/ as in *cat*; final /l/ as in *coal*) and two clusters (monomorphemic final /ld/ as in *cold*; bimorphemic final stop consonant plus plural morpheme /s/ as in *cats*). The final set of target items included 16 words, four with each of these four final consonant patterns, to include in this study (see Table 1).

Ten individual productions of each of the 16 target words were included in the speech-rating task, yielding 160 total stimulus items. The first author judged all of the tokens to be sufficiently noise-free to use in the experiment. All tokens chosen had no perceptible errors or distortions in the onset consonant(s) or the vowel. The final consonants or consonant clusters varied in how closely they resembled MAE-speaking adults' productions of the words. The first author selected stimuli whose final consonants/consonant clusters she judged to vary from fully realized to fully reduced or deleted. Specifically, targets were chosen whose final consonants featured complete or partial cluster reduction (i.e., production of stop + plural /s/ words without /s/), deletion of the final stop consonant, or deletion or vocalization of the final /l/. Some of these variants are typical for AAE (i.e., the production of a stop + plural /s/ without the /s/; Poplack & Tagliamonte, 1994), and some are typical of AAE and other dialects (i.e., the vocalization of /l/ or the glottalization of final /t/, which occurs in many nonstandard regional varieties of American English, as well as in AAE) or casual speech styles (i.e., the production of /ld/ words as /l/; Thomas, 2007). In the remainder of this article, these productions are called *inaccurate* for consistency's sake. The reader is encouraged to keep in mind that they are inaccurate only from the standpoint of MAE production patterns. Each one of them is correct in its own dialect.

The first author's selection of stimuli was blind to an experienced phonetician's transcriptions of the words. On average, 20% of the stimuli selected were transcribed as being produced with a fully absent final consonant. Speech samples from 96 of the 109 talkers were included in the final study, with either one or two words from each selected individual. The stimuli were peak normalized for amplitude across all items and presented to listeners at a comfortable listening level of approximately 70 dB.

Table 1. Target words used in the speech-rating task.

Singleton codas		Cluster codas	
/t/	/l/	/ld/	/t/ + plural /s/
bat	bell	bald	bats
cat	coal	build	books
coat	hole	cold	hands
hat	wheel	gold	hats

In the AV rating task, each talker was paired with a photograph of either an African American or European American child of early elementary age. This was intended to suggest the talker's race to the listeners. The pictures came from an online stock photo subscription site (http://www.superstock.com) and from a collection of licensed images (Eyewire Images, 2002). Because perceived speaker age has been shown to affect listeners' judgments of speech accuracy (e.g., Drager, 2011; Munson et al., 2010), a pilot task was conducted where five respondents estimated the ages of the children in the photographs. No significant difference in perceived age was found between the African American and European American children in the photographs.

The gender of the child paired with the speakers was determined by the first author's judgment of the gender that listeners were likely to identify from that voice. For the 96 talkers, 52 were paired with pictures of girls (26 African American and 26 European American) and 44 with pictures of boys (22 African American and 22 European American). For those speakers from whom two productions were selected, the same picture was paired with both samples of the individual's speech. There were two versions of the AV speech-rating task, such that every speech sample paired with an African American child's face in the first version was paired with a European American child's face in the second and vice versa. Equal numbers of African American and European American children's pictures were paired with each of the four stimulus types (/l/, /t/, /ld/, stop + plural /s/). The same set of images appeared in both versions.

Procedure

The speech-rating tasks were programmed and executed in E-Prime (Version 1.2; Schneider, Eschman, & Zuccolotto, 2002). This took place inside a sound-treated booth for all of the untrained listeners and all but seven of the trained listeners. As seven of the professional SLPs in the trained listener group were unable to come to the lab to complete the study, these individuals were tested on a laptop in a quiet room in their homes or workplaces.

Listeners read the instructions at their own pace on the computer monitor before beginning the task. The instructions informed the participants that they would be completing a study about the accuracy of children's speech production and that they would be seeing pictures of the children who produced the samples. On each trial, the text "Listen to the child say the word [WORD]" was displayed on the screen. For the AV rating task, the computer displayed the picture of a child while the speech sample was delivered over the headphones. For the baseline rating task, the sample was played without a visual image. Listeners

were instructed to rate the accuracy of the speech production on a VAS ranging from *completely accurate* on the left end point to *completely inaccurate* on the right end point. The VAS is shown in Figure 1. For each trial, listeners indicated their perception of the child's speech accuracy by using the mouse to select the corresponding location along this continuum. Listeners were encouraged in the instructions to use the entire line in making their ratings, rather than simply selecting between the two end points.

Participants completed four practice trials after reading the task directions and before beginning the task proper. The practice items consisted of words not targeted in the current study, which were paired with faces of Asian American children in the AV rating task and with no picture in the baseline rating task. On every trial, the computer recorded the coordinates, in pixels, of the listeners' selections for later analysis.

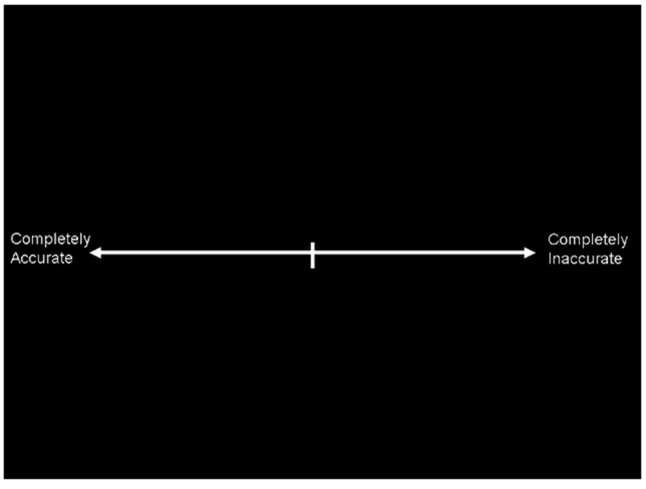
Recall that there were two versions of the AV speech-rating task, such that every speech sample paired with an African American child's face in the first version was paired with a European American child's face in the second and vice versa. Within both the trained and untrained listener groups, half of the participants completed Version 1 and half completed Version 2, though the presentation order of individual test items was randomized for each participant.

Implicit Association Task

Stimuli

The IAT was adapted from that of Babel (2012). This particular IAT uses 20 stereotypically African American names (e.g., Tyrone), 20 stereotypically European American names (e.g., Luke), and two other sets of 20 words, associated with good and bad, respectively. The good words, such as *vacation,* carry generally positive associations, whereas the bad words, like *vomit,* carry negative associations. Babel's complete set of stimulus items were drawn from Greenwald et al. (1998), Dasgupta and Greenwald (2001), and Jelenec and Steffens (2002). Because Dasgupta et al. (2000) previously established that familiarity with the names used in the IAT did not affect performance, this factor was not controlled for in this study. A longer description of this IAT, including the full list of words and names, can be found in Babel (2012).

Figure 1. The visual analog scale used in the speech-rating task. The horizontal line was 444 pixels long and centered 309 pixels from the left side of the screen.



Procedure

Like the speech-rating task, the implicit association task was programmed and executed in E-Prime (Schneider et al., 2002). Participants completed this task immediately following the speech-rating task, in the same environment and using the same computer equipment as in the previous task. The experimenter delivered instructions to the participants orally following the completion of the speech-rating task. The "1" and "3" numeral keys on the keyboard were assigned to correspond to the category choices displayed on the left-hand and right-hand sides of the computer monitor, respectively. The experimenter instructed each participant to use only the index finger of his or her dominant hand to select between these two keys.

The task comprised five blocks. To prevent fatigue, participants were allowed to pause between each task block. During the first, second, and fourth blocks, the two concepts (Black and White, Blocks 1 and 4) or attributes (good and bad, Block 2) were displayed in the upper corners of the monitor, one per side. Randomly selected names (Blocks 1 and 4) or words (Block 2) were then presented in the center of the screen, and participants were instructed to press the button indicating the associated concept or attribute as quickly and accurately as possible. After each response, the word *correct* or *incorrect* appeared in the center of the screen. The use of feedback is consistent with the methods that have been used previously for this IAT and for many other related IATs. During the test blocks (Blocks 3 and 5), the concepts and the attributes were both displayed in the upper corners of the screen, one above the other. Randomly selected names (to be categorized by concept, ignoring the attributes) and words (to be categorized by attribute, ignoring the concepts) appeared in the center of the screen. Figure 2 shows an example of the display during one of these test blocks. The two test blocks differed in which concept was paired visually with which attribute. For each item, the computer recorded both the answer choice and response latency, which was later analyzed to determine each individual's implicit association score, related to the difference in response latency between the two test blocks (see Results section).

Participant Questionnaires

A set of questionnaires was designed to elicit measures of participants' explicit knowledge of the features of AAE, their explicit attitudes relating to AAE, and, for the trained listener group, their experience working with

Figure 2. Example of the display from Block 3 of the Implicit Association Test.

AAE-speaking clients. The AAE Knowledge Survey and the Language Attitude Survey were programmed in E-Prime, so participants completed these parts of the tasks immediately following completion of the speech-rating task. They selected their responses by pressing the letter (AAE Knowledge Survey) or number (Language Attitude Survey) corresponding to each answer choice on the keyboard. The remainder of the questionnaire was completed in paper-and-pencil format following the completion of the implicit association task. This portion included demographic information and questions for the trained listeners about their years of experience and the estimated percentage of their caseloads made up of speakers of AAE.

The Language Attitudes Survey took the form of a 25-item survey consisting of statements such as "AAE is lazy English" and "AAE would be inadequate for teaching subjects such as social studies or math." Respondents indicated their level of agreement with each statement according to a 7-point Likert-type scale. The survey items were drawn from language attitude measures used by Vafadar and Utt (1992) and Blake and Cutler (2003), which appeared originally in Hoover, McNair-Knox, Lewis, and Politzer (1996).

To assess the participants' explicit knowledge of the phonological and morphosyntactic features of AAE, all listeners completed a 22-item, multiple-choice quiz (the AAE Knowledge Survey). In some quiz items, participants chose the pair of words, from a field of three pairs, which would sound alike when produced by a speaker of AAE or southern English. In other items, participants chose the phrase, wording, or expression most typical of AAE or southern English or interpreted the meaning of AAE expressions. The questions on this quiz originally appeared in Ford et al. (1975).

Results

Implicit Association Task

The response latencies for each response made during the five blocks of the IAT test were recorded by the computer and analyzed to determine each individual's IAT score. This study used the revised IAT scoring procedures outlined in Greenwald, Nosek, and Banaji (2003). The resulting score, *d,* compares the difference in response latencies between the target attribute block (where "Black" was paired visually with "good") and the reversed target attribute block (where "White" was paired with "good"). A positive score indicates a pro–European American bias, a negative score indicates a pro–African American bias, and a score of 0 is neutral. There was not a significant difference between groups, $t(38) = 0.886$, $p = .381$ (experienced: $M = 0.324$, $SD = 0.378$; inexperienced: $M = 0.216$, $SD = 0.389$).

Knowledge Measure

Participants' answers on the 22-item, multiple-choice AAE knowledge quiz were recorded by the computer and scored by assigning 1 point to every correct answer and 0 points to every incorrect answer and calculating the percentage correct for every participant. The mean score for the experienced listeners was 78.4% ($SD = 9.3\%$). The mean for the inexperienced listeners was 72.6% ($SD = 7.8\%$). This difference was significant, $t(38) = 2.154$, $p = .038$). The average score on the 22 items that measured knowledge of phonological variation were used as a predictor of individual differences in the effect of race on ratings of children's speech accuracy. These scores ranged widely for both groups. The trained listeners' scores ranged from 58% to 92%, and the untrained listeners' scores ranged from 58% to 100%. In each group, there was one listener whose scores were significantly greater than chance only at a $p = .067$ level (per the binomial test). That is, there was one trained listener and one untrained listener whose responses were not reliably better than chance at the $\alpha = .05$ level. These listeners were included in the individual differences analyses nonetheless. All other listeners' performances were robustly better than chance.

Explicit Attitudes Measure

A summary score (referred to henceforth as the Explicit Attitude Test or EAT score) for the dialect attitude survey was calculated by summing responses on all items, reversing the scoring on some items so that in all cases, a higher numbered response indicated a position that was anti-AAE, pro-MAE, more prescriptive, or more likely to interpret dialect features as a disorder. Correspondingly, low-numbered responses indicated positions that were pro-AAE, more descriptive in nature, and more likely to interpret dialect features as a language difference. There was a significant difference between groups, $t(38) = 2.604$, $p = .013$; the trained listeners ($M = 141.5$, $SD = 15.05$) had higher EAT scores overall than did the untrained listeners ($M = 127.9$, $SD = 17.86$). Given the diversity of topics measured by this questionnaire, one question, specifically, was chosen as a predictor of individual differences in the effect of imputed race on speech accuracy. This was the question that gauged people's agreement with the statement "AAE is lazy English." The two groups did not differ significantly in their responses to this measure (Wilcoxon W = 218, $p = .62$), though the median score was higher (indicating greater disagreement with the statement) for the trained listeners ($M = 6.5$) than the

untrained listeners (M = 5.0). The trained listeners' scores ranged from 1 (indicating *strongest agreement with the statement*) to 7, and they ranged from 2 to 7 for the untrained listeners.

Speech-Rating Task

Before beginning analyses, the x and y coordinates of all mouse clicks on the speech-rating task were examined for outliers. When participants clicked beyond the end points of the line, located 87 and 530 pixels from the left side of the screen, those values were rounded up or down to 87 or 530, respectively. All x coordinates were then scaled to be a percentage of the line, such that a click on the midpoint of the inaccurate–accurate continuum had an x value of 50, 0 indicated a click at the "completely inaccurate" end of the scale, and 100 indicated a click at the "completely accurate" end of the scale. These values are shown in the figures in the remainder of this section.

Baseline Task

The first analyses examined the ratings collected in the baseline task. Recall that these were used as continuous measures of the accuracy of children's speech, against which the accuracy measures from the AV task were compared. Hence, the goal of the analysis of these data was to find the optimal summary measure of each item's accuracy in the absence of a visual prime that suggested the speaker's race.

Violin plots of individual items were examined. No item elicited clearly bimodal ratings. However, the ratings for some items were skewed. Hence, we used the median ratings for each item. The distribution of median ratings for the 160 items was itself skewed, such that there were more ratings toward the completely accurate end of the scale. This was expected, given that there were more items transcribed to be correct than transcribed to be incorrect. These values were transformed to resemble a normal distribution, so that the statistical tests that referenced them would be more robust. The values were rescaled so that the lowest value was 1, and then the square root of the rescaled value was taken. The resulting values were much more normally distributed than the original values.

To determine the face validity of these measures, they were submitted to a two-factor, between-subjects analysis of variance (ANOVA) by items, with coda type (four levels: /l/, /ld/, /t/, stop + plural /s/) and transcribed accuracy (two levels: inaccurate, accurate) as between-subjects factors. There were significant effects of coda type, $F(3, 152) = 13.834$, $p < .001$, and transcribed accuracy, $F(1, 152] = 59.128$, $p < .001$, but no interaction. Post hoc Tukey tests showed pairwise differences in accuracy between all coda types except /t/ and /l/. The /ld/ sequences were rated as the least accurate, and the stop + plural /s/ as the most accurate. Importantly, there was a wide range in perceived accuracy levels for sounds that were transcribed as accurate. This is consistent with evidence cited earlier that there are meaningful phonetic differences among words and sounds that have been transcribed identically. All of the listeners in the baseline condition were asked an open-ended question about what the most salient characteristics of the stimuli for the task were. Only one of the listeners mentioned race. Most of the listeners mentioned that the stimuli varied in accuracy. Hence, we conclude that the ratings in the baseline condition were, at most, only minimally affected by perception of the speaker's race.

Audiovisual Rating Task

Linear mixed-effects models (LMERs) were used to examine these data. LMERs have gained popularity recently as an alternative to by-subject and by-item analyses like ANOVA and multiple regression. In LMERs, each response by each subject to each stimulus is a dependent measure in the statistical model. That is, the data are not averaged by participants across items or averaged by items across participants. Overall differences in the dependent measure (in this case, accuracy ratings) for individual subjects and individual items are modeled explicitly. Because of this, LMER has been argued to be superior to ANOVA or regression, as the outcomes of an LMER are unlikely to be spuriously significant because of a small set of items that elicit a particular response, or a small set of participants who behave a particular way.

The R package lme4 was used to fit the data (Version 1.1-9; Bates, Mächler, Bolker, & Walker, 2015), and the package lmerTest was used to evaluate significance (Version 1.1-0; Kuznetsova, Brockhoff, & Christensen, 2017). Each contrast was examined separately. For each contrast, model building began by fitting a base model with only random intercepts for listeners and items. More complex models were built by adding additional fixed factors, one at a time. Whenever a fixed factor was added, we also added a random slope for the effect of that factor on listeners, items, or both, following the recommendations of Barr, Levy, Scheepers, and Tily (2013). A random slope is a coefficient in a model that estimates the effect of a particular manipulation on subjects or items. For example, when we added imputed race to our statistical model, we also estimated the extent to which imputed race affected ratings for a particular item (as every item was paired with both a European American and African American children's picture) and how imputed race affected individual listeners' ratings (as every listener rated items paired both with European American and African American children's faces). When we added a factor to the model, it was retained if the resulting model fit the data significantly better than the model without the factor. For each contrast, models were built for the entire set of listeners to examine the effects of baseline accuracy, imputed race (i.e., whether the word was accompanied by a picture of an African American or a European American child), group (trained vs. untrained), implicit attitudes about AAE (operationally defined as performance on the IAT), attitudes about AAE (operationally defined as responses to the attitudes survey question asking whether the listeners believed AAE is "lazy speech"), and knowledge of AAE phonology (operationally defined as performance on the phonology questions of the AAE

knowledge test) on ratings. For the trained listeners only, an additional two models were built to examine whether years of experience and proportion of AAE-speaking clients affected ratings. Some transformations were made to the data prior to statistical analysis. All of the numeric ratings were centered prior to analysis. Contrast coding was used for the categorical values of group and race. Years of experience were log-transformed, and the proportion of AAE-speaking clients was square-root transformed to improve the normality of these distributions. These transformations are commonly made when using LMER. They generally result in a greater likelihood that a well-fitting model will be found.

For the model predicting ratings of /l/, a model with a fixed effect for baseline ratings fits the data significantly better than the baseline model, $\chi^2(df = 2) = 121.76$, $p < .001$. A model including a fixed effect for imputed race (including an interaction term between imputed race and baseline accuracy) fits the data even better, $\chi^2(df = 8) = 33.093$, $p < .001$. A model including a fixed effect for group (including a three-way Group × Imputed Race × Baseline Accuracy interaction) did not improve model fit at the $\alpha = .05$ level but did approach this, $\chi^2(df = 5) = 10.671$, $p = .058$. Adding fixed effects for knowledge, implicit attitudes, explicit attitudes, and years of experience did not improve model fit. Moreover, years of experience and proportion of AAE clientele did not improve model fit on a model with baseline rating and imputed race for the trained listeners only.

The coefficients for the model, including baseline rating, imputed race, and group, are shown in Table 2. As this table shows, the coefficient for the three-way interaction did not achieve statistical significance using the conventional $\alpha = .05$ level but did approach this level. To explore the reason for this interaction, we plotted the relationship between baseline accuracy and accuracy in the audiovisual experiments separately for the stimuli paired with African American and European American children's faces, and separately by group. This is shown in Figure 3. As this figure shows, the trained listeners rated the stimuli as more accurate when paired with African American children's faces than when paired with European American faces. This interacted with baseline accuracy, such that the biggest effect of race on ratings was found for the least accurate tokens; there was no difference in the median ratings for the most accurate tokens. In contrast, the ratings for the untrained listeners did not differ as a function of race.

The next set of models examined perception of the words with final /t/. The same model-building scheme was used. A model including a fixed effect for baseline ratings fits the data significantly better than one with only random slopes for subjects and items, $\chi^2(df = 2) = 257.63$, $p < .001$. A model that included a fixed effect for imputed race (including an interaction with baseline accuracy) did not improve model fit, $\chi^2(df = 8) = 12.36$, $p = .136$. However, a fully factorial model with baseline accuracy, imputed race, and group did fit the data better than a model with only baseline accuracy, $\chi^2(df = 5) = 28.97$, $p < .001$. Adding fixed effects for knowledge, implicit attitudes, explicit attitudes, and years of experience did not improve model fit. Moreover, years of experience and proportion of AAE clientele did not improve model fit on a model with baseline rating and imputed race for the trained listeners only.

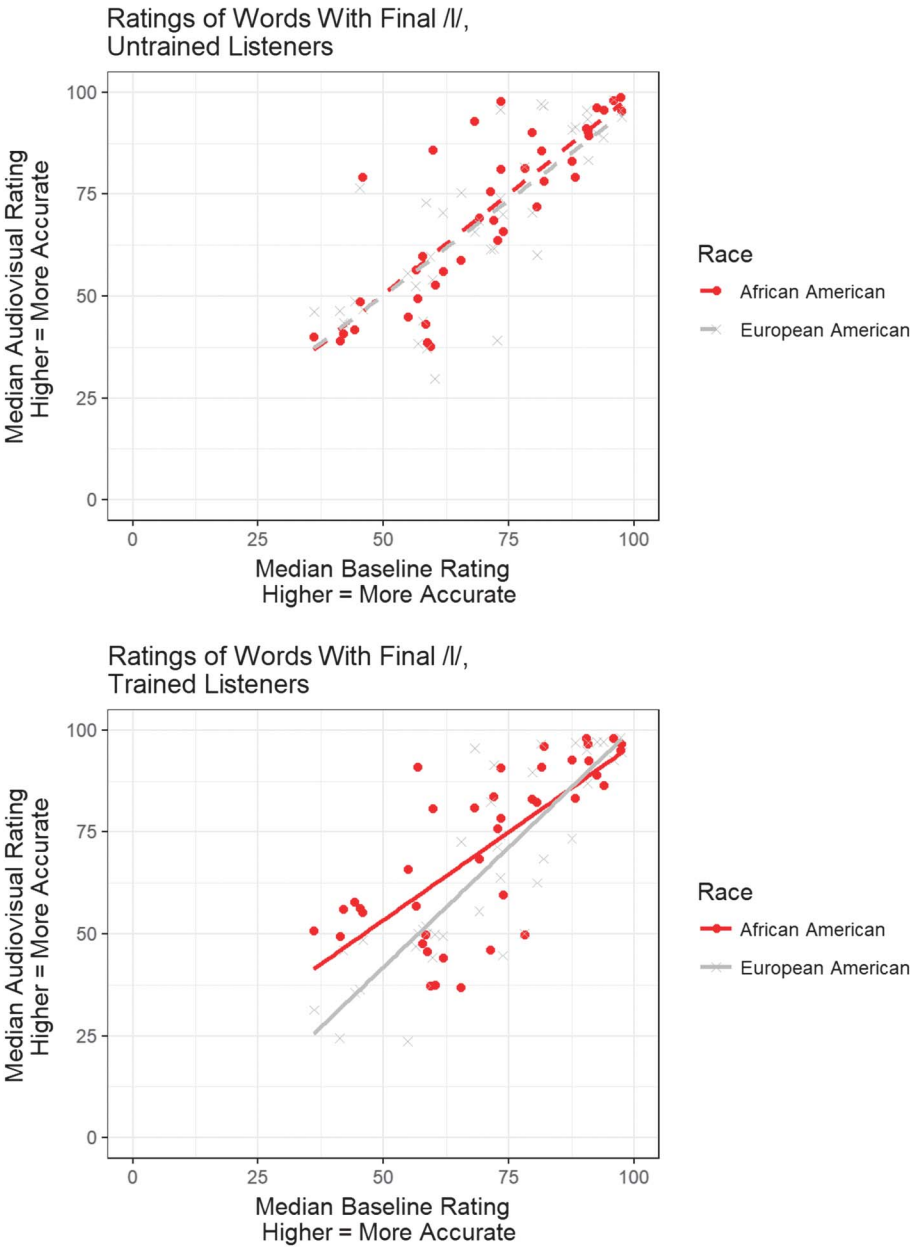
The coefficients for the model, including baseline rating, imputed race, and group, are shown in Table 3. As this table shows, there was, surprisingly, no three-way Baseline Accuracy × Imputed Race × Group interaction, despite this model having significantly better fit than simpler models with only two of these three factors. This may be due to the increase in model fit being driven by the random effects, for which there are no conventional significance tests. Moreover, the coefficient for imputed race did not achieve statistical significance using the conventional $\alpha = .05$ level, but did approach this level. Figure 4 plots the data in a manner parallel to that of Figure 3. As this Figure shows, the ratings for words paired with African American faces were higher than those paired with European American faces, and the difference between ratings was larger for words that were judged by the baseline listeners to be less accurate. However, the magnitude of the difference was similar for both groups. The only difference between the groups is that the median ratings were higher for the trained listeners than for the untrained listeners.

The next set of models examined words ending with /ld/. A model including a fixed effect for baseline ratings fits the data significantly better than one with only random slopes for subjects and items, $\chi^2(df = 2) = 292.8$, $p < .001$. A model including a fixed effect for imputed race (including an interaction term between imputed race and baseline accuracy) resulted in a model that did not converge. Removing

Table 2. Coefficients for the most complex model predicting the accuracy of final /l/ words.

Factor	Estimate	*SE*	*df*	*t* value	*p* value
(Intercept)	66.738	2.115	52.0	31.54	< .001
Baseline accuracy	0.745	0.083	56.5	8.92	< .001
Imputed race	2.548	0.659	36.1	3.86	< .001
Group	–0.286	1.946	40.1	–0.14	.883
Baseline Accuracy × Imputed Race	–0.083	0.043	30.4	–1.90	.066
Baseline Accuracy × Group	0.003	0.068	40.1	0.05	.954
Imputed Race × Group	0.799	0.552	409.6	1.44	.148
Baseline Accuracy × Imputed Race × Group	–0.073	0.038	37.3	–1.93	.060

Figure 3. Median ratings for the 40 stimuli that ended in /l/, separated by whether they were presented with an African American child's face (red circles, red regression line) or a European American child's face (gray crosses, gray regression line). Untrained listeners are plotted on the top figure (dashed regression lines), and trained listeners (solid regression lines) are plotted on the bottom.



the random slope for the effect of baseline accuracy on individual listeners' ratings resulted in a model that did converge. That model improved upon the fit of a simpler model with a fixed effect for baseline accuracy but not random effect of baseline accuracy on individual listeners, $\chi^2(df = 4) = 15.881$, $p = .003$. None of the models with other factors (group, attitudes toward AAE, implicit attitudes toward African Americans, knowledge of AAE) improved the model fit significantly for the entire group of 40 listeners. Moreover, models with years of experience and proportion of AAE-speaking caseloads did not improve upon models with baseline accuracy and imputed race for the 20 trained listeners.

The coefficients for the model with imputed race and baseline accuracy are shown in Table 4. As this model shows, the interaction between imputed race and baseline accuracy did not achieve statistical significance. This can be seen in Figure 5. The regression line predicting median ratings of stimuli paired with African American children's

Table 3. Coefficients for the most complex model predicting the accuracy of final /t/ words.

Factor	Estimate	*SE*	*df*	*t* value	*p* value
(Intercept)	64.536	2.323	47.8	27.78	< .001
Baseline accuracy	0.912	0.093	55.0	9.76	< .001
Imputed race	1.151	0.632	27.5	1.82	.079
Group	–1.994	2.272	44.6	–0.87	.385
Baseline Accuracy × Imputed Race	–0.031	0.032	32.0	–0.98	.330
Baseline Accuracy × Group	0.079	0.089	49.9	0.89	.377
Imputed Race × Group	0.063	0.626	40.3	0.10	.919
Baseline Accuracy × Imputed Race × Group	–0.004	0.031	71.7	–0.15	.878

Figure 4. Median ratings for the 40 stimuli that ended in /t/, separated by whether they were presented with an African American child's face (red circles, red regression line) or a European American child's face (gray crosses, gray regression line). Untrained listeners are plotted on the top figure (dashed regression lines), and trained listeners (solid regression lines) are plotted on the bottom.

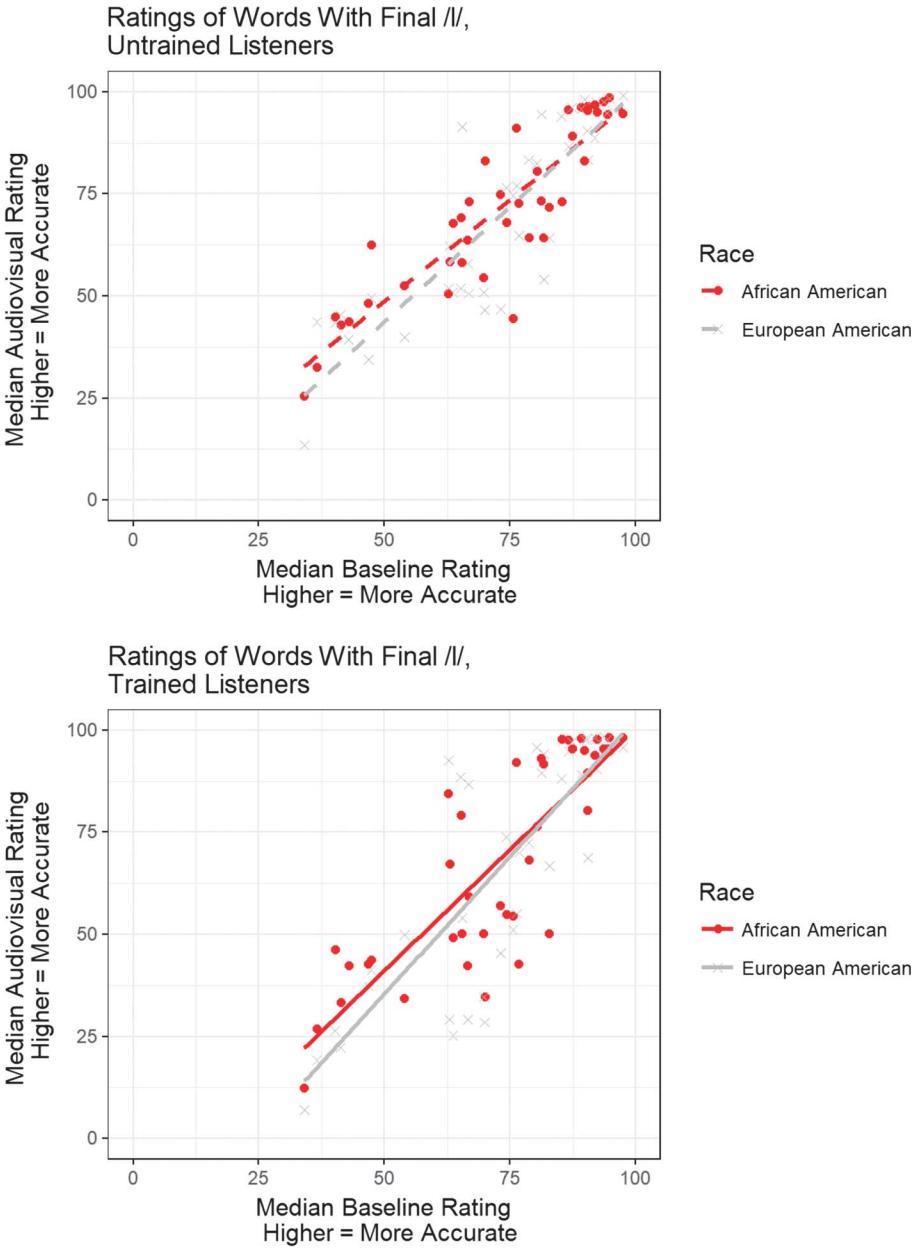


Table 4. Coefficients for the most complex model predicting the accuracy of final /ld/ words.

Factor	Estimate	*SE*	*df*	*t* value	*p* value
(Intercept)	68.536	2.051	50.8	33.422	< .001
Baseline accuracy	0.763	0.038	37.6	19.892	< .001
Imputed race	1.583	0.776	30.1	2.040	.050
Baseline Accuracy × Imputed Race	−0.040	0.027	36.9	−1.485	.146

faces from baseline ratings is higher than the line predicting ratings of stimuli paired with European American children's faces, but the lines are parallel.

The final set of models examined words ending with stop + plural /s/. A model including a fixed effect for baseline ratings fits the data significantly better than one with only random slopes for subjects and items, $\chi^2(df = 2) = 215.43$, $p < .001$. A model including a fixed effect for imputed race (including an interaction term between imputed race and baseline accuracy) resulted in an improvement in model fit, $\chi^2(df = 8) = 23.026$, $p = .003$. None of the more complex models fit the data, either for the entire group of listeners or for the group of trained listeners only. The coefficients for the most complex model are shown in Table 5, and the data are plotted in Figure 6. The plot in Figure 6 clearly illustrates the significant interaction between baseline accuracy and imputed race from Table 5. The biggest difference between the regression lines for stimuli paired with African American children's faces and those for European American children's faces are for the stimuli that have the lowest baseline accuracy values. The lines converge for the stimuli with the highest baseline accuracy ratings.

Discussion

The main finding in this work is that both clinically trained and untrained listeners rated children's speech as more accurate when paired with African American faces than European American faces, a manipulation we call *imputed race*. This is true across three of the four stimulus word types examined: final /l/, final /ld/, and final stop + plural /s/, for which there were statistically significant effects of imputed race on ratings. For stimuli with final /t/, the *p* value for the influence of imputed race was .079. An examination of the coefficients for imputed race shows the strongest effects for the final /l/ stimuli and for the final stop + plural /s/ stimuli. The fact that the stop + plural /s/ stimuli were the subject to a strong effect of race was not surprising. Of the four features that we examined, this one is arguably the most unique to AAE. It is more surprising that imputed race influenced ratings of words with final /l/ strongly. While it is true that some studies have reported a higher incidence of vocalized /l/ variants in AAE speakers, even more studies have shown this to be characteristic of many regional varieties of American English whose speakers

Figure 5. Median ratings for all 40 listeners' ratings of the 40 stimuli that ended in /ld/, separated by whether they were presented with an African American child's face (red circles, red regression line) or a European American child's face (gray crosses, gray regression line).

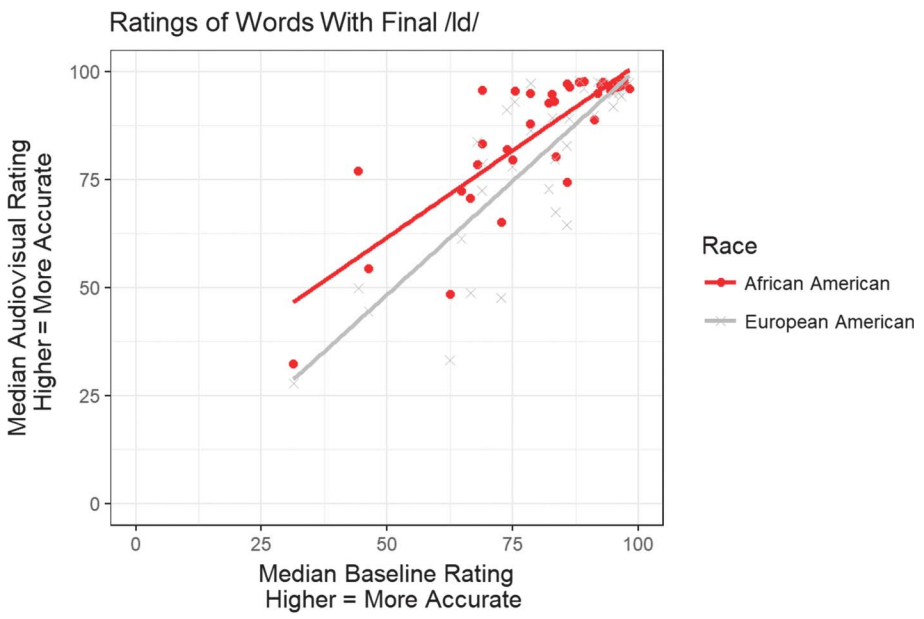


Table 5. Coefficients for the most complex model predicting the accuracy of final /t/ + plural /s/ words.

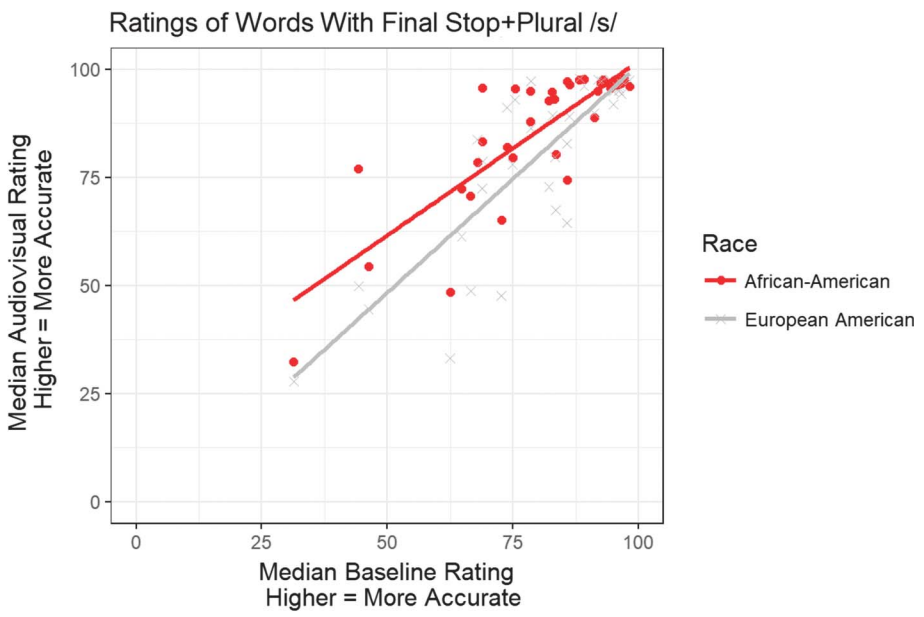
Factor	Estimate	*SE*	*df*	*t* value	*p* value
(Intercept)	69.896	2.460	60.1	28.411	< .001
Baseline accuracy	–6.456	0.854	65.5	–7.555	< .001
Imputed race	2.536	0.613	40.5	4.132	< .001
Baseline Accuracy × Imputed Race	0.746	0.272	70.6	2.742	.008

are racially diverse. For example, Ash (1982) showed that /l/ vocalization is pervasive among European American speakers of the regional variant of American English spoken in Philadelphia. It is unclear why words with /l/ are particularly susceptible to biasing from imputed race. Conversely, it is less surprising that imputed race has the smallest influence on ratings of words with final /t/. Phonetic variation in final stop consonants, particularly final /t/, is also pervasive in the United States and is reported to be a feature of AAE. In this sense, it parallels the patterns of variation associated with /l/. There is no obvious explanation for why /l/-final and stop-final words patterned differently in this study. This might simply reflect the possibility that the variation in final /l/ is inherently more perceptually salient than is the variation in final stops, given that /l/ is a longer and more intense sound than is a final stop. One methodological innovation in this study is that we conducted a baseline experiment, in which a group of listeners rated words' accuracy without suggesting the race of the children who produced them using a continuous rating scale. This baseline experiment verified that the stimuli varied continuously in their perceived accuracy.

One factor that motivated this study was to examine whether African American children would be incorrectly penalized for productions that are accurate in AAE. Our results suggest the opposite, namely, that listeners are less likely to mislabel a production as inaccurate when the speaker is African American. This is consistent with previous work on adults' perception of AAE by Staum Casasanto (2008).

One surprising finding in this study was that clinical training was not consistently associated with differences in ratings. We reasoned that listeners with clinical training would be more likely than untrained listeners to accommodate for presumed AAE use when speech tokens were paired with African American children's faces. We did not find this to be so. This may be evidence that exposure to AAE through social interactions and media portrayals may be sufficient to learn the pronunciation patterns that were examined in this study. Another surprising finding was that measures of knowledge of and attitudes about AAE did not predict the effect of imputed race on ratings. We reasoned that listeners with negative implicit and explicit attitudes toward African Americans and toward AAE would

Figure 6. Median ratings for all 40 listeners' ratings of the 40 stimuli that ended in stop + plural /s/, separated by whether they were presented with an African American child's face (red circles, red regression line) or a European American child's face (gray crosses, gray regression line).



show the greatest effect of imputed race on ratings by most strongly penalizing tokens paired with African American children's faces. The reasons for these findings are unclear. One possibility is simply the general nature of these measures in assessing bias. An instrument that is more narrowly focused on attitudes toward and knowledge of AAE might be needed to show individual differences.

There are at least three areas of future research that this study suggests. The first of these should explore why the ratings for /l/-final words were most consistently affected by speakers' imputed race. Perhaps, this is because vocalized or deleted final /l/ varies more consistently as a function of speaker ethnicity than does any of the other three variants we studied. This seems unlikely, given how pervasive final /l/ variation is in different nonstandard varieties of American English. A more likely explanation is that overtly held stereotypes about AAE are more strongly associated with the variation in final /l/ than in the other three variables we examined. This could be examined empirically.

A second area of research should examine a problem that is faced not only by this study but also by Staum Casasanto and by many other studies of social influences in speech perception. Specifically, research must determine whether viewing an African American or European American truly activates linguistic stereotypes. Given that not all African Americans speak AAE and that some European Americans do speak AAE, there is likely to be substantial individual variation in individuals' association between ethnicity and dialect use. A listener whose European American and African American acquaintances use many AAE features would be predicted to have a weaker effect of imputed race on ratings in this study than would one who only hears AAE from African Americans. These individual differences would be logistically challenging to study but have the potential to help explain the individual differences we observed in this study. The failure of the individual-differences measures in this study to predict differences in the magnitude of imputed race effects highlights the need to develop better individual-differences measures for the study of AAE.

Finally, the findings in this study should be replicated with different stimuli. The importance of replication is twofold. First and foremost, replication is the basis for sound theory building. Second, replication is particularly important for studies of the type presented in this article, where a strictly social expectation is thought to have influenced behavior. Many seminal findings on social priming on behaviors other than speech perception have recently been questioned, as some of the key findings in that literature have failed to be replicable (Shanks et al., 2013). These findings place an especially strong onus on researchers to demonstrate that findings like those in this article are indeed robust across listeners, stimuli, and tasks. Moreover, a replication of this study could also include an extension to one, including common developmental errors that are not related to AAE. Including those errors could test one alternative hypothesis that listeners have a general bias to rate speech more positively when it is paired with African American children's faces, perhaps as a form of stereotype suppression (Wyer, Sherman, & Stroessner, 1998).

These results are relevant to the field of clinical speech-language pathology in that they enhance our nascent understanding of the many ways listener bias and known or assumed speaker characteristics affect speech perception and its clinical consequences. If they were to generalize to clinical settings, then we would expect that equivalent word forms would be rated differently if they were produced by African American or European American children. Though this study did not uncover specific relationships between predictor variables and perceptual ratings, we hope that it sets the stage for a continuing evaluation of these effects. Ultimately, the hope for this line of research is to aid in the development of effective and efficient clinician training, where awareness, discussion, self-assessment, and targeted interventions would serve to reduce bias and support the provision of appropriate and fair services for all children.

Acknowledgments

This work was supported by a Wisconsin Institutes for Discovery Seed grant to Mark Seidenberg; by National Science Foundation Grant BCS0729140 to Jan Edwards; by National Institutes of Health Grant R01 DC02932 to Jan Edwards, Mary E. Beckman, and Benjamin Munson; and by National Institute of Child Health & Human Development Grant P30 HD03352 to the Waisman Center at the University of Wisconsin. Portions of this work were conducted as part of the first author's 2012 master's degree from the University of Minnesota. The authors thank Mark DeRuiter for his input on that document. The authors are very grateful to all of the adult participants and to the children who contributed the speech samples and their families. The authors also thank Mary E. Beckman for useful input on this work, Nicole Breunig who did the original transcriptions of the stimuli, both Hannah Julien and Veera Vasandani for comments on this article, and both Mandi Proue and Carol-June Leonard for valuable assistance in participant recruitment and testing.

References

American Speech-Language-Hearing Association. (2004). *Knowledge and skills needed by speech-language pathologists and audiologists to provide culturally and linguistically appropriate services* [Knowledge and skills]. Retrieved from http://www.asha.org/policy

Ash, S. (1982). *Vocalization of /l/ in Philadelphia*. (Doctoral dissertation). University of Pennsylvania, Philadelphia, PA.

Babel, M. E. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics, 40,* 177–189.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Blake, R., & Cutler, C. (2003). AAE and variation in teachers' attitudes: A question of school philosophy? *Linguistics and Education, 14,* 163–194.

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. (2011). Speaker age and vowel perception. *Language and Speech, 54,* 99–121.

Edwards, J., Gross, M., Chen, J., MacDonald, M., Kaplan, D., Brown, M., & Seidenberg, M. (2014). Dialect awareness and lexical comprehension of mainstream American English in African American English–speaking children. *Journal of Speech, Language, and Hearing Research, 57,* 1883–1895.

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

Ford, J., Lewis, S., Hicks, S., Williams, D., Hoover, M., Politzer, R., & McNair, K. (1975). Tests of African American English for teachers of bidialectical students. In R. Jones (Ed.), *Handbook of tests and measures for Black populations* (Vol. 1, pp. 367–381). Hampton, VA: Cobb & Henry.

Gaither, S., Cohen-Goldberg, A., Gidney, C., & Maddox, K. (2015). Sounding Black or White: Priming identity and biracial speech. *Frontiers in Psychology, 6,* 457.

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., Nolan, A., & Drager, K. (2006). From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review, 23*(3), 351–379.

Holliday, J. J., Reidy, P., Beckman, M. E., & Edwards, J. (2015). Quantifying the robustness of the English sibilant fricative contrast in children. *Journal of Speech, Language, and Hearing Research, 58,* 622–637.

Hoover, M., McNair-Knox, F., Lewis, S., & Politzer, R. (1996). African American English attitude measures for teachers. In R. Jones (Ed.), *Handbook of tests and measurements for Black populations* (Vol. 1, pp. 83–93). Hampton, VA: Cobb & Henry.

Jelenec, P., & Steffens, M. C. (2002). Implicit attitudes toward elderly women and men. *Current Research in Social Psychology, 7*(16), 275–292.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122,* 148–203.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lakoff, R. (2000). *The language wars*. Berkeley, CA: University of California Press.

Lass, N. J., Tecca, J., Mancuso, M., & Black, W. (1979). The effect of phonetic complexity on speaker race and sex identifications. *Journal of Phonetics, 7,* 105–118.

Lindblom, B. (1990). Explaining variation: A sketch of the H and H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 403–439). Dordrecht, the Netherlands: Kluwer.

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

Miller, J. (1981). Effects of speaking rate on segmental distinctions. In P. Eimas & J. Miller (Eds.), *Perspectives on the study of speech* (pp. 39–74). Hillsdale, NJ: Erlbaum.

Munson, B., Edwards, J., Schellinger, S., Beckman, M., & Meyer, M. (2010). Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of *Vox Humana*. *Clinical Linguistics & 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., & Lunders, E. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory and Cognition, 30,* 583–593.

Perkall, J., & Klatt, D. (Eds.). (1986). *Invariance and variability in speech processes*. London, United Kingdom: Psychology Press.

Poplack, S., & Tagliamonte, S. (1994). -S or nothing: Marking the plural in the African-American diaspora. *American Speech, 69,* 227–259.

Rickford, J., & Rickford, R. (2000). *Spoken soul: The story of Black English*. New York, NY: Wiley.

Romeo, R., Hazan, V., & Pettinato, M. (2013). Developmental and gender-related trends of intra-talker variability in consonant production. *The Journal of the Acoustical Society of America, 134,* 3781–3792.

Rubin, D. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education, 33*(4), 511–531.

Schellinger, S. K., Munson, B., & Edwards, J. (2017). Gradient perception of children's productions of /s/ and /θ/: A comparative study of rating methods. *Clinical Linguistics & Phonetics, 31,* 80–103.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime user's guide [Computer program]. Pittsburgh, PA: Psychology Software Tools.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., . . . Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS ONE, 8,* e56515.

Staum Casasanto, L. (2008). *Experimental investigations of sociolinguistic knowledge*. (Doctoral dissertation). Stanford University, Stanford, CA.

Strand, E. A., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (Ed.), *Natural language processing and speech technology: Results of the 3rd KONVENS Conference, Bielefeld, October 1996* (pp. 14–26). Berlin, Germany: de Gruyter.

Thomas, E. (2007). Phonological and phonetic characteristics of African American vernacular English. *Language and Linguistics Compass, 1,* 450–475.

Thomas, E., & Reaser, J. (2004). Delimiting perceptual cues used for the ethnic labeling of African American and European American voices. *Journal of Sociolinguistics, 8,* 54–87.

U.S. Census Bureau. (2012, May 17). *Most children younger than age 1 are minorities, Census Bureau reports* [Press release CB12-90]. Retrieved from https://www.census.gov/newsroom/releases/archives/population/cb12-90.html

U.S. Department of Education. (2001). *Twenty-third annual report to Congress on the implementation of the Individuals with Disabilities Education Act*. Retrieved from http://www2.ed.gov/about/reports/annual/osep/2001/index.html

Vafadar, A., & Utt, H. (1992). A survey of speech-language-pathologists' attitudes about, self-perceived knowledge of, and competency in dealing with social dialects: Language differences versus language disorders. *National Student Speech Language Hearing Association Journal, 20,* 65–72.

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

Wyer, N., Sherman, J., & Stroessner, S. (1998). The spontaneous suppression of racial stereotypes. *Social Cognition, 16,* 340–352.