Attention, Perception, & Psychophysics (2022) 84:2074–2086
https://doi.org/10.3758/s13414-021-02423-w

REGISTERED REPORTS AND REPLICATIONS

Revisiting the relationship between implicit racial bias and audiovisual benefit for nonnative-accented speech

Drew J. McLaughlin[1] · Violet A. Brown[1] · Sita Carraturo[1] · Kristin J. Van Engen[1]

Accepted: 30 November 2021 /Published online: 5 January 2022

Abstract

Speech intelligibility is improved when the listener can see the talker in addition to hearing their voice. Notably, though, previous work has suggested that this "audiovisual benefit" for nonnative (i.e., foreign-accented) speech is smaller than the benefit for native speech, an effect that may be partially accounted for by listeners' implicit racial biases (Yi et al., 2013, *The Journal of the Acoustical Society of America, 134*[5], EL387–EL393.). In the present study, we sought to replicate these findings in a significantly larger sample of online participants. In a direct replication of Yi et al. (Experiment 1), we found that audiovisual benefit was indeed smaller for nonnative-accented relative to native-accented speech. However, our results did not support the conclusion that implicit racial biases, as measured with two types of implicit association tasks, were related to these differences in audiovisual benefit for native and nonnative speech. In a second experiment, we addressed a potential confound in the experimental design; to ensure that the difference in audiovisual benefit was caused by a difference in accent rather than a difference in overall intelligibility, we reversed the overall difficulty of each accent condition by presenting them at different signal-to-noise ratios. Even when native speech was presented at a much more difficult intelligibility level than nonnative speech, audiovisual benefit for nonnative speech remained poorer. In light of these findings, we discuss alternative explanations of reduced audiovisual benefit for nonnative speech, as well as methodological considerations for future work examining the intersection of social, cognitive, and linguistic processes.

Keywords Speech perception · Psycholinguistics

Understanding spoken language requires listeners to process a highly variable acoustic signal. This variability exists both within talkers (e.g., the same phoneme is produced differently in different contexts as a result of coarticulation) and between talkers. One source of between-talker variability that listeners frequently encounter is systematic deviation from native language norms when speech is produced by a nonnative speaker. Not only do listeners show poorer intelligibility for nonnative-accented relative to native-accented speech (Clarke & Garrett, 2004), they also show increased listening effort as indicated by both slower response times to a secondary task and increased pupil dilation (Brown et al., 2020; McLaughlin & Van Engen, 2020).

One cue that listeners use to facilitate speech processing is the visual information provided by the speaker's face. A large body of research has indicated that being able to see as well as hear the talker improves speech intelligibility for syllables (Sommers et al., 2005), words (Erber, 1969; Sumby & Pollack, 1954), and sentences (Tye-Murray et al., 2016; Van Engen et al., 2017)—for both young and older adults (Tye-Murray et al., 2016), for individuals with hearing loss (Tye-Murray et al., 2007), and for cochlear implant users (Kaiser et al., 2003). Indeed, this "audiovisual benefit" is one of the most robust findings in the speech perception literature. However, few studies have addressed the extent to which audiovisual benefit differs for nonnative-accented and native-accented speech, and the mechanisms underlying those differences. Existing evidence suggests that listeners may gain less visual benefit for nonnative-accented speech than for native-accented speech (Babel & Mellesmoen, 2019; Waddington et al., 2020; Xie et al., 2014; Yi et al., 2013).

✉ Drew J. McLaughlin
drewjmclaughlin@wustl.edu

[1] Department of Psychological & Brain Sciences, Washington University in St. Louis, One Brookings Dr, St. Louis, MO 63130, USA

In addition to finding reduced audiovisual benefit for Korean-accented English relative to native-accented American English (for native English-speaking listeners), Yi et al. (2013) also found that this difference in benefit was related to participants' implicit biases. This finding suggested that implicit biases—specifically associations between East Asian faces and the construct *foreign*—may partially impact the efficiency of audiovisual integration for nonnative speech. In a follow-up study that used the same methods but collected fMRI data (Yi et al., 2014), the authors also found that participants' implicit biases were associated with increased BOLD response to audiovisual nonnative speech. Increased activity related to implicit bias was found in the right primary auditory cortex, which is thought to be responsible for early processing of acoustic information (Peelle, 2012; Poeppel, 2003). Measures of implicit bias did not explain activity in the left primary auditory cortex or the left inferior frontal gyrus (associated with language comprehension and cognitive control processing; Goghari & MacDonald 3rd., 2009; Peelle, 2012; Poeppel, 2003). Together, findings from these studies indicated that reduced audiovisual benefit for nonnative speech may be partially explained by listeners' implicit biases.

Implicit association tests (IATs) have been widely used and validated in the field of social psychology for examining the strengths of implicit associations between constructs (Greenwald et al., 1998; Nosek & Smyth, 2007). The outcome in the IAT is response latency, and the basic assumption of the task is that if two constructs are strongly associated, then categorization will be easier when those constructs are paired (e.g., for a Bush supporter, pictures of George Bush, and words synonymous with Good; Greenwald et al., 2003). A wide range of constructs can be examined with IATs, but designs typically examine associations of two sets of contrasted constructs (examining biases toward a single construct, such as with a go/no-go task, tends to result in lower reliability; Nosek & Banaji, 2001). For example, in Yi et al.' (2013) study, they included an IAT that measured associations between White versus East Asian and *American* versus *Foreign*. While often referred to as a measure of implicit bias, it is important to note that IAT captures simultaneous associations between multiple constructs. Thus, the IAT used by Yi and colleagues did not simply capture bias against East Asians. Rather, it captured the strength of associations between East Asian faces and the construct *foreign*, and—simultaneously—associations between White faces and the construct *American*.

Although IATs are widely used to assess implicit associations, there are multiple well-documented limitations and criticisms of the measure. Internal consistencies for IATs are typically fairly high (about 0.70 to 0.90; Nosek & Smyth, 2007), but test–retest reliability is often poor (about 0.50 on average; Lane et al., 2007). A meta-analysis by Greenwald et al. (2009) indicated good predictive validity of IATs for behavioral measures across multiple domains. However, the average effect size for predicting behavioral outcomes with IATs was $r = .27$ across 122 studies. When limiting this summary to IATs specifically examining race, this average was slightly lower (closer to $r = .24$).

Thus, with this meta-analysis in mind, we would expect that the relationship between individual differences in implicit bias and audiovisual integration of speech (if it exists) would be small. The discrepancy between this assessment and the findings of Yi et al. (2013)—which had a sample size of $n = 19$ and found an effect size of $r = .48$—motivated the present replication study.

In Experiment 1, we conducted a direct replication of the two primary findings of Yi et al. (2013). Namely, we sought to replicate the reduced audiovisual benefit found for nonnative- compared with native-accented speech, as well as the relationship between implicit biases and the magnitude of this reduction in audiovisual benefit. For the latter finding, Yi and colleagues found that listeners with stronger implicit associations between East Asian faces and the construct *foreign*, and between White faces with the construct *American*, had reduced audiovisual benefit for nonnative-accented speech (relative to that for native-accented speech). In addition to the measure of implicit *American* versus *Foreign* associations, we added a measure of implicit *Good* versus *Bad* associations. This allowed us to investigate whether the implicit foreignness associations related to audiovisual benefit are dissociable from negative (*Bad*) associations. For example, stereotypes related to cultural foreignness (e.g., having a foreign accent or not speaking English very well) typically distinguish Americans' attitudes toward Asian Americans versus White Americans, while stereotypes related to criminality (e.g., drug abuse) do not (Zou & Cheryan, 2017). Given that the present study examines the effect of implicit biases on the perception of Korean-accented speakers, we thought that this distinction may prove theoretically important. Specifically, we predicted that implicit biases related to the construct *foreign* may explain reduced audiovisual integration for nonnative speech, while implicit biases related to the construct *Bad* may not. In Experiment 2, we aimed to test the robustness of the difference in audiovisual benefit for native versus nonnative accent (discussed further after Experiment 1).

Experiment 1

Method

The preregistered hypotheses and analysis plan for Experiment 1 can be found at https://osf.io/8j9wq. Data and

analysis scripts for both experiments can be found at https://osf.io/wv624/files/.

Participants The sample size ($N = 260$) for the present study was determined via power analysis a priori with a focus on ensuring sufficient statistical power to test the relationship between IAT *d* scores and audiovisual benefit. In the field of social psychology, research that has examined relationships between implicit bias and behavioral measures has typically found relatively small effect sizes (see Greenwald et al., 2009, for a meta-analysis). Although Yi et al. (2013) found an effect size of $r = .48$, we estimated our sample size using an effect size estimate of $r = .20$. Using the *pwr.r.test()* function in R (Version 4.0.4), we determined that in order to obtain 90% power to detect an effect size of $r = .20$ we would need a sample size of approximately 260 participants.

A total of 291 participants were recruited from the Washington University Psychological & Brain Sciences Subject Pool. To match the eligibility criteria from Yi et al. (2013), we required all participants to be native monolingual speakers of English with no known language or hearing issues. Eleven participants were excluded from the sample because they reported being bilingual. Additionally, because we collected data online, we required participants to use headphones to reduce variability in audio quality. Six participants were excluded from the sample because they reported using computer speakers instead of headphones. An additional four participants were excluded because they did not pass the attention-check trials (details below) or they performed three standard deviations below the average performance level in the speech transcription task. Lastly, we included one question at the end of the experiment that asked participants whether there was any reason their data should be excluded. Participants were encouraged to answer honestly, and told that answering "Yes (my data needs to be excluded)" would not affect their participation credit. Ten participants were excluded based on their self-reports. Many of these participants reported in an optional written-response that their data should be excluded because they were too distracted during the task. In total, 31 participants were excluded for not meeting eligibility criteria or for failing to meet our criteria for ensuring data quality.

Information about participants' race/ethnicity and gender was collected in open-response questions. Of the 260 participants retained in the sample, 73 self-reported that their gender was man (or responded "male"), 185 woman (or responded "female"), and two nonbinary. Hispanic/Latinx is an ethnicity that can co-occur with a variety of races; however, multiple subjects reported simply that they were Hispanic and/or Latinx with no further information. Thus, here, we summarize race and ethnicity together: 25 participants reported that they were Asian (includes East, Southeast, and South Asian responses), 33 Black or African American, six Hispanic/Latinx, 170 White, 25 mixed race, and one participant declined to respond. Participants provided informed consent and were compensated with course credit, as approved by the Washington University in St. Louis Internal Review Board. The study lasted approximately 30 minutes.

Materials For the replication, Yi et al. (2013) shared their original audiovisual and audio-only stimuli from their speech transcription task, as well as the images from their *American* versus *Foreign* implicit association task. These stimuli were unaltered for the present experiment, but we report how they were prepared for the original study below for reference.

Speech transcription task. Materials included 40 simple, meaningful sentences containing four keywords each (e.g., "the girl loved the sweet coffee"; Van Engen et al., 2012). Two native-accented American English and two nonnative-accented English speakers produced the targets (one male, one female per accent). The native-accented speakers were both White, and the nonnative-accented speakers were both Korean (with Korean as their L1). Neither of the White talkers spoke with highly salient regional markers. The male White talker grew up in Houston, Texas, and the female White talker grew up in White Plains, New York.

A six-talker babble track created from 30 simple, meaningful sentences (Bradlow & Alexander, 2007) produced by native-accented American English talkers (three male, three female) was used for noise mixing. Random samples from the track were mixed with the target sentences at a −4 dB signal-to-noise ratio. The babble extended 500 ms before and after the target files. For the audiovisual stimuli, the surrounding 500 ms of the video was a freeze-frame image of the speaker.

We created two additional stimuli for attention-check trials. For these targets, a separate female native speaker of American English spoke the sentences "please type a single G" and "please type a single Q." These files were presented without noise in audio-only format. There was no indication to the participants that these trials were the attention-check trials and not one of the 40 primary trials.

Implicit association Tests (IATs). We used the stimuli from Yi et al. (2013) for testing implicit associations between White versus East Asian and *American* versus *Foreign*. These images included faces of ten East Asian young adults (five male, five female), faces of 10 White young adults (five male, five female), 10 iconic American scenes (e.g., the White House), and 10 non-American foreign scenes (e.g., the Eiffel Tower). For the *Good–Bad* IAT, the same face images were used, but keywords were used instead of pictures for the good–bad dimension. *Good* keywords included: *wonderful, pleasant, glorious, nice*, and

Table 1 For each block of the Implicit Association Tests (IATs), the number of trials, the categories, and the function are shown

Block	Number of trials	Left-key (d) response items	Right-key (k) response items	Function
1	20	White	East Asian	Practice: Learn race dimension
2	20	American	Foreign	Practice: Learn attribute dimension
3	20	White + American	East Asian + Foreign	Race-attribute pairing 1 (Analyzed)
4	40	White + American	East Asian + Foreign	Race-attribute pairing 1 (Analyzed)
5	20	East Asian	White	Practice: Relearn race dimension
6	20	East Asian + American	White + Foreign	Race-attribute pairing 2 (Analyzed)
7	40	East Asian + American	White + Foreign	Race-attribute pairing 2 (Analyzed)

Note. The block order above is counterbalanced across participants. For half of the participants, Blocks 1, 3, and 4 (race-attribute pairing 1) are swapped with Blocks 5, 6, and 7 (race-attribute pairing 2). This same procedure was also used for the White versus East Asian and Good versus Bad implicit association test

superior. Bad keywords included: *terrible, horrible, evil, awful*, and *inferior*.

Questionnaire. The questionnaire assessed language experience, age, gender, race/ethnicity, hearing status, and use of headphones versus computer speakers. A final question asked participants to report if there was any reason that their data should be excluded from the experiment.

Procedure All participation occurred online. The experiment was built and delivered using Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Participants completed the tasks in the following order: speech transcription task, *American–Foreign* IAT, *Good–Bad* IAT, questionnaire, debriefing. Procedures matched those described by Yi et al. (2013) as closely as possible, with the exception of the *Good–Bad* IAT, which was a novel addition.[1]

Speech transcription task. Forty target sentences were presented in noise in randomized order. Each target was presented in a single modality by a single speaker and was not repeated (combination of target and speaker/modality was counterbalanced). Thus, from trial-to-trial, participants randomly alternated between audio-only and audiovisual trials, and between the four talkers. For audio-only trials, a fixation cross was presented onscreen. After each target finished, a response box appeared for participants to type their response. The two attention-check trials were fixed to appear at Trials 13 and 27 (i.e., spaced one-third and two-thirds of the way through the task). Responses were scored for accuracy (by keyword) using the open-source tool Autoscore (Borrie et al., 2019) in R (Version 4.0.4). Attention-check trials were removed from the dataset after determining which participants needed to be excluded.

Implicit association tests (IATs). Procedures for the IATs matched standard recommended protocols (Nosek et al., 2005) and scoring guidelines (Greenwald et al., 2003). The IAT is a response time sorting task containing seven blocks, alternating which constructs are sorted together in each block (see Table 1). In other words, it is not straightforwardly a measure of biases against East Asians—rather, it simultaneously measures associations between two sets of contrasted constructs.

During each trial of the IAT, participants are shown a single image or keyword, which they have to quickly sort into one of two categories. The categories change each block, and are always labeled in the left and right upper corners of the screen. In the present study, responses for the left category were made with the 'd' key and responses for the right category were made with the 'k' key. If subjects sort an item into the wrong category, a red 'X' appears in the center of the screen. Additionally, if subjects do not respond within four seconds, the trial will time-out and the next stimulus will be presented.

For calculating *d* scores following Greenwald et al.'s (2003) guidelines, data is analyzed from Blocks 3, 4, 6, and 7 (all other blocks are for practice; see Table 1). First, trials with latencies greater than 10,000 ms are excluded, and participants with more than 10% of trials with latencies less than 300 ms are removed (no participants met this elimination criterion in the present study). The mean latency for correct trials is calculated for each block, and error trial latencies are replaced with these values plus an additional 600 ms. One pooled standard deviation is computed for all trials in Blocks 3 and 6, and another for all trials in Blocks 4 and 7. The average latencies for each block are used to calculate differences between Blocks 3 and 6 and between Blocks 4 and 7 (specifically, later blocks minus earlier blocks). These differences are then divided by their respective pooled standard deviations, and then averaged to obtain a final *d* value. Blocks 3 and 6 are paired together during this

[1] The accent rating task (a separate experiment/set of participants) used in Yi et al. (2013) was not included in this replication.

process because they contain the earlier trials of each race-attribute pairing (see Table 1), which typically have slower response times than Blocks 4 and 7. Both group-wide and individual *d* scores were calculated.

Results

Speech transcription task As in Yi et al. (2013), we used generalized linear mixed-effects regression to model the transcription data with a logit link function. Given that each sentence has four key words, allowing for multiple successes and failures per trial, transcription accuracy was coded as grouped binomial data. In other words, the models predicted both a vector of successes (number of correct words in a given sentence) and a vector of failures (number of incorrect words in a given sentence).

We employed the maximal random effects structure justified by the study design (Barr et al., 2013). Participants and items (sentences) were included as random intercepts. Given that Modality and Accent are manipulated within-subjects and within-items, we also modeled by-participant random slopes for Modality and Accent, and by-item random slopes for Modality and Accent. Supplemental Table 1A contains a summary of the model containing only lower order fixed effects, Supplemental Table 1B contains a summary of the model containing all lower order effects and two-way interactions, and Supplemental Table 1C contains a summary of the full model containing all lower order effects, two-way interactions, and the three-way interaction.

Likelihood ratio tests were used to assess the significance of the contribution of each fixed effect and interaction to the model. The effects of Modality, $\chi^2(1) = 84.29$, $p < .001$, and Accent, $\chi^2(1) = 37.72$, $p < .001$, both significantly improved model fit. Model estimates indicated that performance was better in the audiovisual compared the audio-only condition ($\beta = 1.04$), and worse for the nonnative compared to the native accent condition ($\beta = -1.43$). The interaction between Modality and Accent also significantly improved model fit, $\chi^2(1) = 12.55$, $p < .001$, and indicated that there was greater audiovisual benefit in the native than the nonnative accent condition ($\beta = -0.17$; see Fig. 1).

We deviated from Yi and colleagues' approach to examine individual differences in implicit biases and their relationship with audiovisual benefit (although, see Native Boost formula section, below). In our analysis, we test this relationship within the mixed-effects model described above. Notably, the three-way interaction between Modality, Accent, and *American–Foreign d* score is key for determining whether implicit bias predicts the difference in audiovisual benefit for native versus nonnative accent.

First, we tested whether adding the fixed effect of *d* score improved model fit, and found a nonsignificant result, $\chi^2(1) = 3.34$, $p = .07$ (see Fig. 2). Neither of the interactions between *American–Foreign d* score and modality, $\chi^2(1) = 2.19$, $p = .14$, or accent, $\chi^2(1) = 0.75$, $p = .39$, significantly improved model fit, nor did the three-way interaction between *d* score, modality, and accent, $\chi^2(1) = 2.62$, $p = .11$.

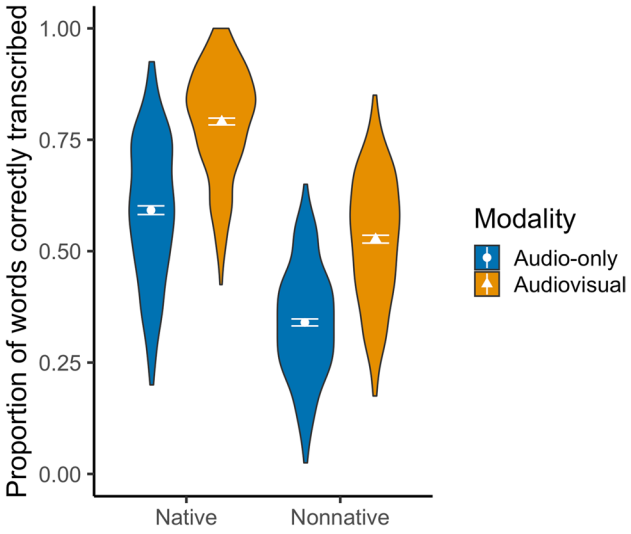

Fig. 1 Performance in Experiment 1, summarized as the proportion of keywords correctly transcribed by-participant, is shown for the two accents (native and nonnative) and two modalities (audio-only and audiovisual). Violin plots display the distribution of individual participants' performance averages, and points show the group means with standard errors

Our preregistered analysis plan specified that if we did not find a significant effect from the three-way interaction then we would test a separate model of the nonnative accent conditions only. This model did not reveal any significant contributions of *American–Foreign d* score, $\chi^2(1) = 0.17$, $p = .68$, or the interaction of *d* score and modality, $\chi^2(1) = 1.52$, $p = .22$, for predicting performance in the nonnative accent condition.

Direct comparison of means across studies. In Table 2, we report the summary statistics of the speech transcription task from Yi et al. (2013) and the current replication. Direct comparison of the raw means reveals that performance in Experiment 1 of the current study was lower than in the original experiment.

Audiovisual benefit formula. For comparison with other work in the field, we also calculated audiovisual benefit for each accent condition using the following formula (Grant et al., 1998; Sommers et al., 2005): (AV − AO) / (1 − AO). This value was calculated for each individual in each condition, and then averaged across participants for each condition. This audiovisual benefit formula standardizes the benefit of adding the visual signal by dividing it by the amount a participant could possibly improve from seeing the talker. The average standardized audiovisual benefit was 42% in

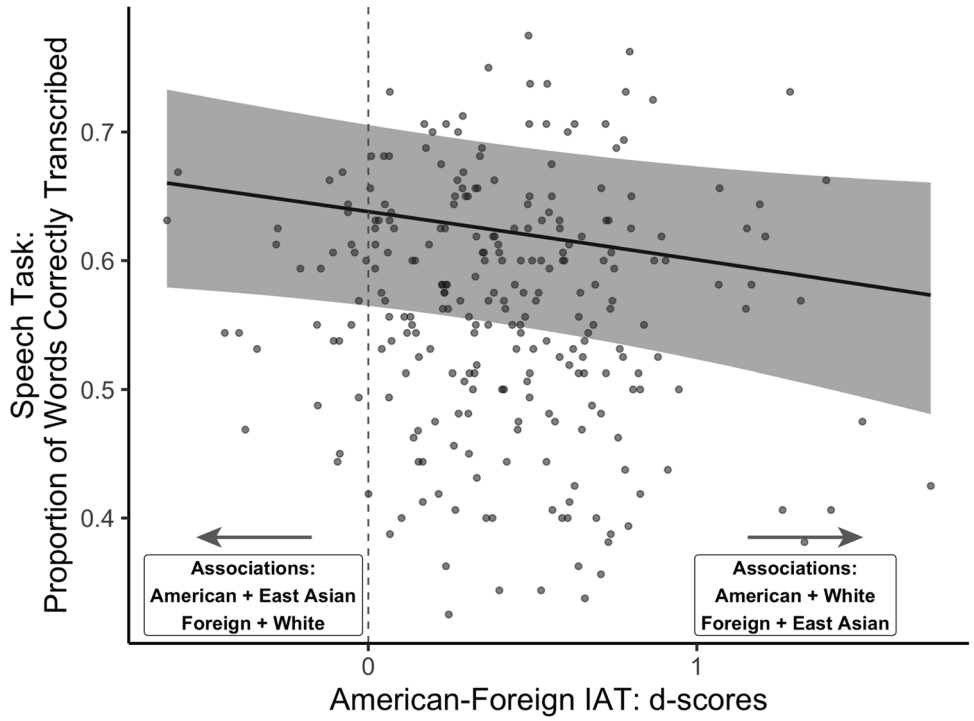

Fig. 2 The slight (but nonsignificant) negative relationship between performance on the speech transcription task and the *American–Foreign* IAT *d* scores is shown with individual subject points and a model fit with 95% confidence interval ribbon. Values above zero on the IAT indicate stronger associations between *American* and White, and between *Foreign* and East Asian (values below zero indicate the opposite)

Table 2 Summary statistics show percentages of keywords correctly identified by modality and accent for Yi et al. (2013) versus Experiment 1 of the current study

Condition	Yi et al. (2013)	Experiment 1	Raw difference
Native audio-only	62.4%	59.1%	3.3%
Native audiovisual	92.9%	79.1%	13.8%
Nonnative audio-only	39.5%	34.0%	5.5%
Nonnative audio-visual	62.5%	52.7%	9.8%

the native condition and 27% in the nonnative condition. A linear model was used in R to compare individual benefit scores and confirmed that audiovisual benefit was lower in the nonnative condition as compared to the native condition (β = -0.15, $p < .001$).

Native boost formula. For comparison with the Yi et al. (2013) study, we also conducted correlation analyses of native boost scores and *American–Foreign d* scores. Native boost was calculated separately for the audio-only and audiovisual modalities, using Yi and colleagues' formula: (native − nonnative) / (1 − nonnative). For the audio-only conditions, native boost scores had a small negative relationship with *American–Foreign d* scores ($r = -0.13$, $p = .04$, CI = [−0.25, −0.01]), and for the audiovisual conditions there was no significant correlation between native boost and *American–Foreign d* scores ($r = 0.04$, $p = .53$, CI = [−0.08, 0.16]). Notably, Yi and colleagues found the opposite pattern of results, such that the audio-only native boost scores showed no relationship with *American–Foreign d* scores, and the audiovisual native boost showed a significant relationship with *American–Foreign d* scores (discussed further below).

Good–Bad IAT. Likelihood ratio tests were also used to test whether *Good–Bad d* scores (instead of *American–Foreign d* scores) were related to performance on the speech transcription task and audiovisual benefit. These models were exactly the same as those described above for the analysis of the *American–Foreign d* scores. The fixed effect of *Good–Bad d* scores did not improve model fit, $\chi^2(1) = 0.24$, $p = .63$. Further, neither the interaction of *Good–Bad d* scores with modality, $\chi^2(1) = 1.02$, $p = .31$, nor with accent, $\chi^2(1) = 1.88$, $p = .17$, improved model fit. Lastly, the three-way interaction between *Good–Bad d* scores, modality, and accent did not improve fit, $\chi^2(1) = 1.83$, $p = .18$.

Implicit association tests (IATs) Linear mixed-effects models were used to examine the effects of condition (congruent, incongruent) and counterbalance order on response time in

the *American–Foreign* and *Good–Bad* IATs. Condition, in the case of the IATs, refers to the race-attribute pairings; the "congruent" condition paired [*American* + White] and [*Foreign* + East Asian] (or [*Good* + White] and [*Bad* + East Asian]), and the "incongruent" condition paired the reverse. Counterbalance order refers to whether subjects first completed the congruent or incongruent blocks. Participants were included as random intercepts with random slopes by condition. Supplemental Table 2A and Supplemental Table 2B summarize the linear mixed-effects models of the lower order and higher order terms from the *American-Foreign* IAT, and Supplemental Table 3A and Supplemental Table 3B summarize the lower order and higher order terms from the *Good-Bad* IAT, respectively.

For the *American–Foreign* IAT, likelihood ratio tests indicated that the effect of condition significantly improved model fit, $\chi^2(1) = 89.32$, $p < .001$, and the effect of counterbalance order did not, $\chi^2(1) = 1.51$, $p = .22$. The model estimate for the condition effect ($\beta = 101.77$) indicated that response times were slower when participants sorted [White + *Foreign*] and [East Asian + *American*] than when they sorted [White + *American*] and [East Asian + *Foreign*]. The interaction between condition and counterbalance order was significant, $\chi^2(1) = 13.10$, $p < .001$, and the direction of the estimate ($\beta = -71.25$) indicated that participants in the second counterbalance order had a smaller effect of condition than those in the first counterbalance order. This effect of counterbalancing is commonly seen in IATs (Teige-Mocigemba et al., 2016), because participants have difficulty switching how they pair races and attributes (i.e., after Block 4; see Table 1).

For the *Good–Bad* IAT, condition also significantly improved fit, $\chi^2(1) = 26.95$, $p < .001$, but counterbalance order did not, $\chi^2(1) = 3.10$, $p = .08$. The estimate for the condition effect ($\beta = 52.50$) indicated slower response times when participants sorted [White + *Bad*] and [East Asian + *Good*] than when they sorted [White + *Good*] and [East Asian + *Bad*]. The interaction between condition and counterbalance order was not significant, $\chi^2(1) = 1.02$, $p = .31$.

Group-wide *d* scores were also calculated for each of the IATs. In the *American–Foreign* IAT, the group-wide statistic was positive ($d = .41$), indicating a moderate bias towards [White + *American*] and [East Asian + *Foreign*]. The group-wide statistic for the *Good–Bad* IAT was near zero ($d = -0.06$). However, upon further inspection, it became clear that two participants had *d* scores greater than three standard deviations below the mean, which were pulling this value down. After removing these two outliers, the group-wide statistic was positive ($d = .20$), indicating a bias towards [White + *Good*] and [East Asian + *Bad*]. Notably, the calculation without the outliers better matches the outcomes of the linear mixed-effects regression, which uses an analysis approach that is typically more robust against outliers.

Lastly, we calculated the Pearson correlation of the individual *d* scores from each of the IATs using the *cor.test()* function in R ($r = .24$, $p < .001$, CI = [.15, .38]). The trends in the individual and group-wide *d* scores for both the *American–Foreign* and *Good–Bad* IAT are shown in Fig. 3. We also calculated Cronbach's alpha for each IAT and found moderate internal consistency (*American–Foreign:* $\alpha = 0.44$; *Good–Bad:* $\alpha = 0.70$).

Exploratory analyses. Given the diversity of our sample, we decided to explore whether implicit biases varied based on participant race (Devos & Banaji, 2005). Specifically, we expected that White participants would have stronger biases than participants of other races, as measured by both IATs. We created a variable labeling participants as White ($n = 170$) or other race ($n = 90$) for this analysis. One notable limitation of this approach is that it grouped together participants with varying racial identities. As such, this analysis primarily examined the effect of White-ness on implicit racial biases. We determined that this was the best option available given that there were not enough participants of each racial category to examine trends on a finer scale.

The effect of participant race was added to the mixed-effects models prior to testing the interaction with condition, though it did not significantly improve model fit for either the *American–Foreign*, $\chi^2(1) = 0.003$, $p = .96$, or the *Good–Bad*, $\chi^2(1) = 0.05$, $p = .83$, IAT. For the *American–Foreign* IAT, likelihood ratio tests indicated that participant race did not significantly interact with condition, $\chi^2(1) = 0.28$, $p = .60$. However, for the *Good–Bad* IAT, the interaction with condition was significant, $\chi^2(1) = 14.63$, $p < .001$, and the direction of the interaction ($\beta = 78.01$) indicated that White participants had slower response times than did participants of other races when sorting together [White + *Bad*] and [East Asian + *Good*]. In other words, White participants had a stronger [White + *Good*] and [East Asian + *Bad*] associations than did non-White participants.

Our a priori plan was to explore the effect of participant race within each IAT, and if these IATs were related to the speech task, we would then examine subsets of the data from the speech task based on participant race. However, given that participant race only affected IAT scores for *Good–Bad* associations, and the *Good–Bad* IAT showed no relationship with the speech transcription task, we did not explore the effect of participant race any further.

Interim discussion

In Experiment 1, we conducted a direct replication of Yi et al. (2013). Our aim was to replicate the reduced audiovisual benefit found for nonnative-accented as compared with native-accented speech, as well as replicate the relationship between this reduced audiovisual benefit and implicit racial

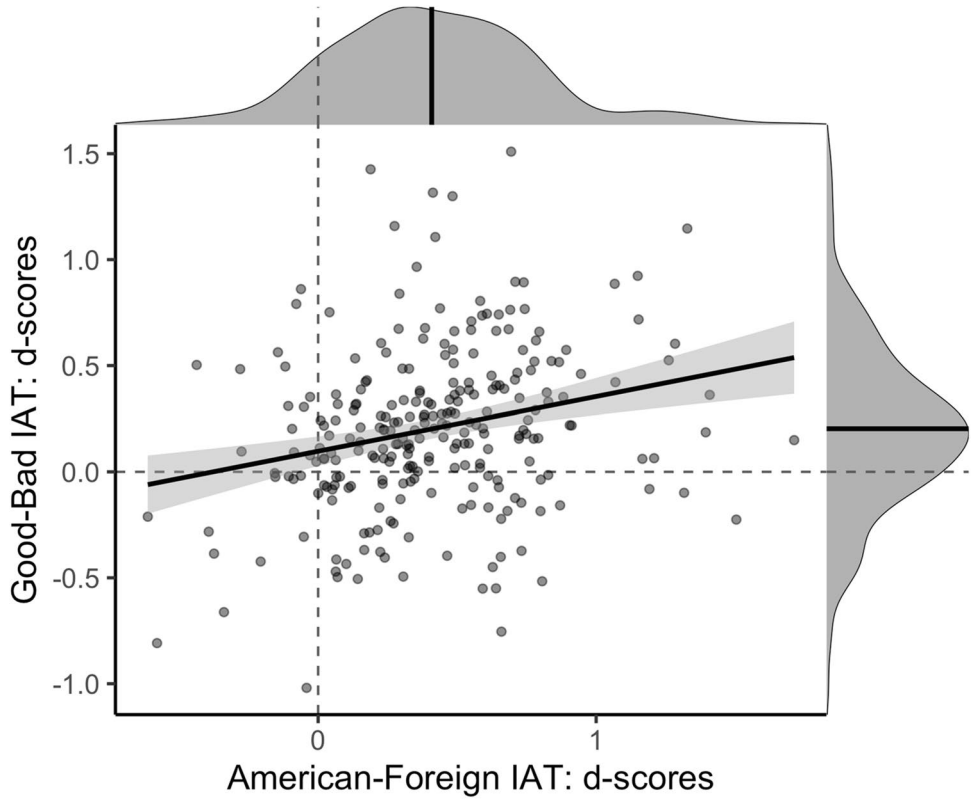

Fig. 3 The correlation between individual *d* scores in each of the implicit association tests (IATs) is shown with scattered points and a best fit line with standard error. Distributions show the spread of individual *d* scores on each task, with a solid line marking the group-wide *d* score (after removing outliers in the *Good-Bad* IAT). A dashed line is included at zero on each axis for reference. For the *American-Foreign* IAT, values greater than zero indicate stronger [White + *American*] and [East Asian + *Foreign*] associations, and for the *Good-Bad* IAT, values greater than zero indicate stronger [White + *Good*] and [East Asian + *Bad*] associations (less than zero indicates the opposite)

biases. We replicated the finding that audiovisual benefit is reduced for nonnative-accented speech, but we did not replicate the finding that implicit racial biases (as measured by an IAT) are related to the magnitude of this effect. We found a small, nonsignificant, negative relationship between implicit racial bias (as measured with a White vs. East Asian and *American* vs. *Foreign* IAT) and overall performance on the speech perception task. However, individual differences in implicit racial biases were not significantly predictive of differences between native versus nonnative accent perception, audio-only versus audiovisual perception, or the interaction between these factors. When using the native boost formula introduced by Yi and colleagues, we found that implicit racial biases had a small negative relationship with native boost, but only for the audio-only conditions and not the audiovisual conditions. Importantly, this finding is the opposite of what Yi and colleagues found (in their study only native boost scores from *audiovisual* conditions were related to implicit biases), and does not support the conclusion that implicit biases explain differences in audiovisual benefit for native versus nonnative talkers. Altogether, the results of Experiment 1 indicate that implicit racial biases are unlikely to be the cause of reduced audiovisual benefit for nonnative speech.

Experiment 2

In our replication (and Yi and colleagues' original study), both native and nonnative speech were presented at the same signal-to-noise ratio (−4 dB SNR), resulting in overall poorer performance in the nonnative compared to the native condition. Prior work indicates a U-shaped relationship between intelligibility of the speech signal and the degree of audiovisual benefit. Typically, there is an increase in audiovisual benefit as the intelligibility of the speech signal is reduced (e.g., from 100% to 50%), but as listening performance drops to extremely low values (e.g., less than approximately 50% accuracy), the visual signal provides less benefit (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). In Experiment 1, the native speakers were 69.1% intelligible, and the nonnative speakers were 43.4%

intelligible (collapsing across modalities). Thus, one alternative explanation for the findings of Experiment 1 is that we observed a smaller audiovisual benefit for nonnative speech than for native speech because the overall difficulty of that condition was greater. In Experiment 2, we aimed to test whether there would still be reduced audiovisual benefit for nonnative speech if the overall difficulties of the two accent conditions were reversed. In other words, we sought to confirm that the observed difference in audiovisual benefit was due to the accent of the speakers, not the difficulty of the task.

Methods

The preregistered hypotheses and analysis plan for Experiment 2 can be found at https://osf.io/6kpuj.

Participants For Experiment 2, paid participants were recruited with Prolific (Palan & Schitter, 2018). The target sample size for Experiment 2 ($N = 110$) was estimated using R (Version 4.0.4) by bootstrapping data from Experiment 1. By simulating models of random samples of participants, we were able to predict the power to detect a significant interaction between accent and modality at multiple sample sizes. The results of this simulation indicated that with approximately 110 participants there would be greater than 80% power to detect a significant interaction between accent and modality.

We recruited a total of 119 participants online, resulting in 110 participants after exclusions. Prescreening settings on Prolific selected for participants residing in the United States, between 18 and 35 years of age, who were monolingual English speakers with no known language or hearing issues. Despite the prescreening, five participants were excluded and replaced for failing to meet the language background criteria. Four additional participants were excluded from the sample because they reported using external speakers instead of headphones, and/or because they reported that their data should be excluded. None of the participants recruited for Experiment 2 failed the attention-check trials.

Race/ethnicity and gender information were collected in open-response questions. Of the 110 participants retained in the sample, 52 reported that their gender was man (or responded "male"), 57 woman (or responded "female"), and one agender. For race/ethnicity, two participants reported that they were Asian, five Black or African American, two Hispanic/Latinx, one Native American, 86 White, 13 mixed-race, and one participant declined to respond. Participants provided informed consent and were compensated at a rate of $10 per hour ($3.33 for less than 20 minutes of participation), as approved by the Washington University in St. Louis Internal Review Board.

Speech transcription task The speech transcription task from Experiment 1 remained exactly the same, with the exception that the native and nonnative speech were presented at signal-to-noise ratios of −5 dB and +3 dB, respectively. Piloting of the stimuli indicated that these new signal-to-noise ratios approximately matched the intelligibility of the audio-only conditions to those in Experiment 1 (i.e., native speakers would be approximately 30% intelligible and nonnative speakers would be approximately 60% intelligible). In other words, these signal-to-noise ratios were selected for the native and nonnative stimuli in order to "flip" the approximate difficulty levels of the accents.

Questionnaire The same questionnaire was used in Experiment 2 as in Experiment 1. Although, because IATs were not included in Experiment 2, the questionnaire immediately followed the speech transcription task.

Results

Model specifications matched those described for Experiment 1. Supplemental Table 4A and Supplemental Table 4B summarize the generalized mixed-effects models, which examined lower order and higher order terms, respectively.

We used likelihood ratio tests to assess whether the contributions of each fixed effect and interaction significantly improved model fit. The effects of modality, $\chi^2(1) = 71.03$, $p < .001$, and accent, $\chi^2(1) = 41.40$, $p < .001$, both improved model fit. As expected, model estimates indicated that performance was better in the audiovisual compared the audio-only condition ($\beta = 1.15$), and better for the nonnative compared with the native accent condition ($\beta = 1.55$). The interaction between modality and accent also significantly improved model fit, $\chi^2(1) = 61.89$, $p < .001$, and indicated that there was greater audiovisual benefit in the native than the nonnative accent condition ($\beta = -0.63$; see Fig. 4), consistent with the results of Experiment 1.

Audiovisual benefit formula As in Experiment 1, we also calculated audiovisual benefit for each accent condition for comparison against other work in the field. This standardized audiovisual benefit was 31% in the native condition and 32% in the nonnative condition. Scores from the nonnative and native condition did not significantly differ, as determined with a linear model ($\beta = 0.01$, $p = .64$).

Exploratory analyses Given that the results from the audiovisual benefit formula deviated from the results of our generalized linear mixed-effects analysis, we decided to conduct post hoc comparisons of audiovisual benefit across experiments. The goal of this exploratory analysis was to determine if the null finding from the audiovisual benefit formula

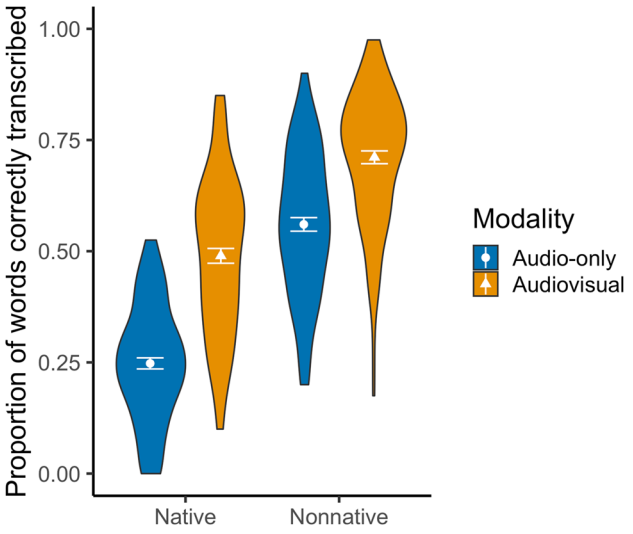

Fig. 4 A summary of the proportion of keywords correctly transcribed by accent (native and nonnative) and modalities (audio-only and audiovisual) is shown for Experiment 2. Violin plots display the distribution of individual participants; performance averages, and points show the group means with standard errors

may be due to the large difference in overall performance between conditions.

First, we compared the audiovisual benefit scores from the native and nonnative conditions of each experiment, which were more closely matched for overall intelligibility. We constructed a model that compared the native condition from Experiment 1 and the nonnative condition from Experiment 2 (the higher intelligibility conditions), and then a model that compared the nonnative condition from Experiment 1 and the native condition from Experiment 2 (the lower intelligibility conditions). The model comparing the high-intelligibility conditions showed less audiovisual benefit for the nonnative condition ($\beta = -.10$, $p = .03$), but the model comparing the low-intelligibility conditions showed no difference between conditions ($\beta = -0.03$, $p = .21$). It is important to note, however, that statistical power for both of these models was poorer than the power for the main analyses because these exploratory analyses examined differences across participants (instead of within subject) with summary statistics.

Next, we used the full datasets and generalized linear mixed-effects models to compare the higher intelligibility conditions and the lower intelligibility conditions (with the expectation that this would improve power to detect differences in audiovisual benefit). In both models, the interaction between accent and modality indicated a significantly smaller audiovisual benefit for the nonnative as compared with native accent ($ps < .001$). Importantly, this indicates that the audiovisual benefit formula may have "overcorrected" for differences between conditions when intelligibility in the audio-only conditions was drastically different.

General discussion

Across two experiments, we replicated the finding that audiovisual benefit for nonnative-accented speech is reduced relative to audiovisual benefit for native-accented speech. However, we did not replicate the second finding from Yi et al. (2013); results of the present study indicated that individual differences in implicit racial biases do not explain reduced audiovisual benefit for nonnative speech. In this discussion, we first review the findings of the study, and then address alternative explanations for reduced audiovisual benefit and directions for future work on the topic.

Our results robustly indicated that listeners garnered more audiovisual benefit for native-accented English than for Korean-accented English. In our first experiment, target stimuli were presented in the same level of noise for both accents, resulting in overall poorer intelligibility of the Korean-accented speakers. Thus, one alternative explanation for our findings was that audiovisual benefit was reduced for the nonnative-accented as compared with the native-accented speech not as a result of the accent per se, but rather the difference in the overall challenge of the listening task. To address this potential confound, in our second experiment we "flipped" our design by adjusting the noise levels of each accent such that intelligibility was poorer overall for the native accent than the for nonnative accent (whereas in Experiment 1, intelligibility was poorer for the nonnative than for the native accent). Our primary regression analysis indicated the same result as in Experiment 1: Audiovisual benefit for nonnative-accented speech was smaller as compared with the benefit for native-accented speech.

We also conducted an analysis of standardized audiovisual benefit, which controls for the "room to improve" across conditions (using the audiovisual benefit formula from Grant et al., 1998; Sommers et al., 2005). Using these standardized values, we found a difference in audiovisual benefit for native versus nonnative accent in Experiment 1, but not in Experiment 2. Notably, when comparing the difficulty-matched conditions across experiments with mixed-effects modeling, we again found reduced audiovisual benefit for the nonnative condition as compared to the native condition. We interpret the null finding with the standardized scores from Experiment 2 as being due to the large difference in overall difficulty between the native and nonnative conditions, which may not be conducive to making comparisons with the standardized audiovisual benefit formula.

When comparing the raw mean performance levels across studies, we found that subjects in Experiment 1 of the

current study performed more poorly than the subjects in Yi et al. (2013). One possible explanation for this difference is that data was collected online in the present study. Although we required that listeners use headphones, there was nonetheless reduced control over the quality of audio presentation and the testing environment of the listener, both of which may have reduced overall performance. It is unlikely that this difference in overall accuracy negatively affected the investigation of differences in audiovisual benefit for native versus nonnative accents. The combined results of Experiments 1 and 2 clearly indicate a difference in audiovisual benefit across accents that persists across levels of listening difficulty.

Another difference between the original study by Yi et al. (2013) and the current replication is the sampled population of participants. Yi and colleagues sampled college students residing in the Austin, Texas area, whereas we sampled college students residing in the St. Louis, Missouri, area. Both the University of Texas at Austin and Washington University in St. Louis recruit students nationally, but it is reasonable to assume that the regional composition of our samples differed. In particular, the original sample likely included a larger proportion of students from Texas. It is possible that the regional varieties of the native-accented White talkers in the study were perceived differently by these groups, although neither talker spoke with highly salient regional markers. However, given the consistent finding of reduced audiovisual benefit for nonnative accent across our studies, we do not believe that participant background systematically affected results.

Contrary to Yi et al. (2013), we did not find evidence of a relationship between implicit racial biases and reduced audiovisual benefit for nonnative speech. In the present study, we used the same materials and procedures, with the exception that data were collected online instead of in-lab. Importantly, we increased the sample size for the experiment drastically, predicting that if there is a relationship between implicit racial bias and audiovisual benefit, the effect size is likely to be small. Even with our larger sample, we found no evidence to support the conclusion that differences in audiovisual benefit for native and nonnative speakers are explained by listeners' implicit racial biases.

Multiple accounts may explain the reduced audiovisual benefit for nonnative-accented relative to native-accented speech. First, because nonnative speech deviates from native productions acoustically, the increased cognitive demands (Brown et al., 2020; McLaughlin & Van Engen, 2020) and/or slower overall rate of processing (Adank et al., 2009) for nonnative speech may negatively affect processes such as audiovisual integration (*acoustic stream account*). Second, it is possible that listeners are less adept at identifying nonnative visemes and/or matching these visemes to their acoustic counterparts (*visual stream account*). Lastly, it may be the case that listeners with greater racial biases engage less with nonnative and/or minority-race speakers, either implicitly or explicitly devoting fewer cognitive resources to processing audiovisual speech (*racial bias account*). Our results do not support this third account of reduced audiovisual benefit for nonnative speech. However, we also acknowledge that the null results of the present study cannot definitively rule out racial biases as a factor that affects audiovisual speech processing. Further, these accounts of reduced audiovisual benefit are not mutually exclusive. It is possible that a combination of the acoustic and visual stream accounts may explain differences in audiovisual benefit for native relative to nonnative speech.

Another important question for future research is whether audiovisual processing of nonnative speech can improve with training and/or experience. A large body of work (primarily using audio-only materials) has indicated that listeners can rapidly improve their ability to understand nonnative-accented speech (i.e., perceptual adaptation; see Baese-Berk, 2018). It is possible that audiovisual integration for nonnative speech also improves with training/experience.

Methodological considerations for future work

Implicit association tests (IATs) are widely used in social psychology to assess implicit biases (see review by Nosek et al., 2007). However, for use with speech perception research, we note an important limitation of the IAT that must be considered for future work. Namely, because IATs often have poor reliability (particularly test–retest reliability; Lane et al., 2007), relatively large sample sizes are needed when using IATs as measures of individual differences. This concern is what motivated the current replication of Yi et al. (2013), in which a relatively small sample ($n = 19$) revealed a moderate relationship ($r = .48$) between IAT scores and differences in audiovisual benefit between native- and nonnative-accented speech. Moving forward, it is crucial that interdisciplinary work combining individual participants' IAT measures (or any measure of individual differences) with linguistic measures considers how the precision of these measures affects statistical power. Indeed, studies with lower power are less likely to replicate (Maxwell, 2004). Power can be improved by increasing either the sample size or the precision of measures (e.g., adding more trials), the latter of which is a helpful option for reducing research costs.

An additional methodological issue emerged during our analyses when using the audiovisual benefit formula (Grant et al., 1998; Sommers et al., 2005). The purpose of the formula is to create a standardized value that summarizes the benefit of adding the visual signal to the auditory signal. Importantly, the audiovisual benefit formula takes into account the amount a participant could possibly improve

from seeing the talker. For example, if intelligibility in the audio-only condition is 60%, then listeners can gain 40% improvement from seeing the talker, but if audio-only intelligibility is 85%, then the maximum benefit a listener can obtain is only 15%. In the present study, we found that the audiovisual benefit formula may actually overcorrect for such differences in "room to improve" between conditions. Indeed, when conducting analyses with mixed-effects models we found robust differences in audiovisual benefit for native- as compared to nonnative-accented speech, regardless of whether intelligibility was better for native-accented speech, nonnative-accented speech, or relatively well-matched across accents. When using the audiovisual benefit formula, however, these differences did not emerge in all cases. In part, the differing outcomes for the two analyses may be attributable to differences in power (i.e., the mixed-effects analysis gains power by examining trial-level rather than aggregated data). Nonetheless, we recommend that future researchers carefully consider the appropriateness of the audiovisual benefit formula for their research designs and analyses.

Conclusion

A large body of work examining audiovisual speech perception indicates that speech intelligibility is improved when the listener can see a talker in addition to hearing their voice. However, prior work has demonstrated that this audiovisual benefit is reduced for nonnative-accented relative to native-accented speech (Babel & Mellesmoen, 2019; Waddington et al., 2020; Yi et al., 2013). The present study replicates this finding, and shows that it is stable across varying levels of speech intelligibility. Nonetheless, we did not find evidence for the claim that this difference in audiovisual benefit is related to listeners' implicit racial biases.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13414-021-02423-w.

Acknowledgments We want to express our gratitude to Han-Gyol Yi, Jasmine E. B. Phelps, Rajka Smiljanic, and Bharath Chandrasekaran for sharing the materials from their original experiment so we could conduct a direct replication. This work was supported by Graduate Research Fellowships (DGE-1745038) awarded to Drew J. McLaughlin and Violet A. Brown by the National Science Foundation, and Washington University in St. Louis.

References

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(2), 520–529. https://doi.org/10.1037/a0013552

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Babel, M., & Mellesmoen, G. (2019). Perceptual adaptation to stereotyped accents in audio-visual speech. *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia* (pp. 1044–1048). http://intro2psycholing.net/ICPhS/papers/ICPhS_1093.pdf. Accessed date 01 Mar 2020

Baese-Berk, M. (2018). Perceptual learning for native and non-native speech. In K. D. Federmeier & L. Sahakyan eds.), *Psychology of learning and motivation* (Vol. 68, pp. 1–29). Elsevier. Accessed date 01 Mar 2020 https://www.sciencedirect.com/science/article/pii/S007974211830001X

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). https://doi.org/10.1016/j.jml.2012.11.001

Borrie, S. A., Barrett, T. S., & Yoho, S. E. (2019). Autoscore: An open-source automated tool for scoring listener perception of speech. *The Journal of the Acoustical Society of America*, *145*(1), 392. https://doi.org/10.1121/1.5087276

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America*, *121*(4), 2339–2349. https://doi.org/10.1121/1.2642103

Brown, V. A., McLaughlin, D. J., Strand, J. F., & Van Engen, K. (2020). Author accepted manuscript: Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/1747021820916726

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658. https://doi.org/10.1121/1.1815131

Devos, T., & Banaji, M. R. (2005). American = White? *Journal of Personality and Social Psychology*, *88*(3), 447. https://psycnet.apa.org/journals/psp/88/3/447.html?uid=2005-01818-003. Accessed date 01 Mar 2020

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, *12*(2), 423–425. https://doi.org/10.1044/jshr.1202.423

Goghari, V. M., & MacDonald, A. W., 3rd. (2009). The neural basis of cognitive control: Response selection and inhibition. *Brain and Cognition*, *71*(2), 72–83. https://doi.org/10.1016/j.bandc.2009.04.004

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America*, *103*(5, Pt. 1), 2677–2690. https://doi.org/10.1121/1.422788

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037//0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*(1), 17–41. https://doi.org/10.1037/a0015575

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing*

Research: JSLHR, *46*(2), 390–404. https://www.ncbi.nlm.nih.gov/pubmed/14700380. Accessed date 01 Mar 2020

Lane, K. A., Banaji, M. R., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the implicit association test: IV: What We know (so far) about the method. In B. Wittenbrink (Ed.), Implicit measures of attitudes (Vol. 294, pp. 59–102). Guilford. https://psycnet.apa.org/fulltext/2007-01388-003.pdf. Accessed date 01 Mar 2020

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

McLaughlin, D. J., & Van Engen, K. J. (2020). Task-evoked pupil response for accurately recognized accented speech. *The Journal of the Acoustical Society of America*, *147*(2), EL151–EL156. https://doi.org/10.1121/10.0000718

Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. *Social Cognition*, *19*(6), 625–666. https://doi.org/10.1521/soco.19.6.625.20886

Nosek, B. A., & Smyth, F. L. (2007). A Multitrait-multimethod validation of the implicit association test. *Experimental Psychology*, *54*(1), 14–29. https://doi.org/10.1027/1618-3169.54.1.14. Accessed date 01 Mar 2020

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the implicit association test: II. Method variables and construct validity. *Personality & Social Psychology Bulletin*, *31*(2), 166–180. https://doi.org/10.1177/0146167204271418

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit association test at age 7: A methodological and conceptual review. Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes, *341*, 265–292. https://psycnet.apa.org/fulltext/2007-00387-006.pdf

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Peelle, J. E. (2012). The hemispheric lateralization of speech processing depends on what "speech" is: A hierarchical perspective. *Frontiers in Human Neuroscience*, *6*, 309. https://doi.org/10.3389/fnhum.2012.00309

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as "asymmetric sampling in time". *Speech Communication*, *41*(1), 245–255. https://doi.org/10.1016/S0167-6393(02)00107-3

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cerebral cortex, 17(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing*, *26*(3), 263–275. https://doi.org/10.1097/00003446-200506000-00003

Sumby, W. H., & Pollack, I. (1954). Visual contributions to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. https://doi.org/10.1121/1.1907309

Teige-Mocigemba, S., Klauer, K. C., & Sherman, J. W. (2016). *A practical guide to implicit association task and related tasks*. https://escholarship.org/content/qt63t6n75d/qt63t6n75d.pdf. Accessed date 01 Mar 2020

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing*, *28*(5), 656–668. https://doi.org/10.1097/AUD.0b013e31812f7185

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., & Sommers, M. S. (2016). Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. *Psychology and Aging*, *31*(4), 380–389. https://doi.org/10.1037/pag0000094

Van Engen, K. J., Chandrasekaran, B., & Smiljanic, R. (2012). Effects of speech clarity on recognition memory for spoken sentences. *PLOS ONE, 7*(9), e43753. https://doi.org/10.1371/journal.pone.0043753

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception & Psychophysics*, *79*(2), 396–403. https://doi.org/10.3758/s13414-016-1238-9

Waddington, E., Jaekel, B. N., Tinnemore, A. R., Gordon-Salant, S., & Goupell, M. J. (2020). Recognition of Accented speech by cochlear-implant listeners: Benefit of audiovisual cues. *Ear and Hearing*, *41*(5), 1236–1250. https://doi.org/10.1097/AUD.0000000000000842

Xie, Z., Yi, H.-G., & Chandrasekaran, B. (2014). Nonnative audiovisual speech perception in noise: Dissociable effects of the speaker and listener. *PLOS ONE*, *9*(12), e114439. https://doi.org/10.1371/journal.pone.0114439

Yi, H.-G., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2013). Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of the Acoustical Society of America*, *134*(5), EL387–EL393. https://doi.org/10.1121/1.4822320

Yi, H.-G., Smiljanic, R., & Chandrasekaran, B. (2014). The neural processing of foreign-accented speech and its relationship to listener bias. *Frontiers in Human Neuroscience*, *8*, 768. https://doi.org/10.3389/fnhum.2014.00768

Zou, L. X., & Cheryan, S. (2017). Two axes of subordination: A new model of racial position. *Journal of Personality and Social Psychology*, *112*(5), 696–717. https://doi.org/10.1037/pspa0000080

Open Practices Statement Both Experiments 1 and 2 are preregistered. Data and materials for both are publicly available (https://osf.io/wv624/files/).