Revisiting the left ear advantage for phonetic cues to talker identification

Lee Drown,[1,2] Betsy Philip,[1] Alexander L. Francis,[3] and Rachel M. Theodore[1,2]

[1]Department of Speech, Language, and Hearing Sciences, University of Connecticut, Storrs, CT, 06269-1085, United States

[2]Connecticut Institute for the Brain and Cognitive Sciences, University of Connecticut, Storrs, CT, 06269-1272, United States

[3]Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, IN, 47907-2122, United States

Author to whom correspondence should be addressed:

Rachel M. Theodore

rachel.theodore@uconn.edu

ABSTRACT

Previous research suggests that learning to use a phonetic property (e.g., voice-onset-time, VOT) for talker identity supports a left ear processing advantage. Specifically, listeners trained to identify two "talkers" who only differed in characteristic VOTs showed faster talker identification for stimuli presented to the left ear compared to the right ear, interpreted as evidence of hemispheric lateralization consistent with task demands. Experiment 1 (n = 97) aimed to replicate this finding and identify predictors of performance; experiment 2 (n = 79) aimed to replicate this finding under conditions that better facilitate observation of laterality effects. Listeners completed a talker identification task during pre-test, training, and post-test phases. Inhibition, category identification, and auditory acuity were also assessed in experiment 1. Listeners learned to use VOT for talker identity, which was positively associated with auditory acuity. Talker identification was not influenced by ear of presentation, with Bayes Factors indicating strong support for the null. These results suggest that talker-specific phonetic variation is not sufficient to induce a left ear advantage for talker identification; together with the extant literature, they instead suggest that hemispheric lateralization for talker-specific phonetic variation requires phonetic variation to be conditioned on talker differences in source characteristics.

Keywords: speech perception; talker identification; hemispheric lateralization; perceptual learning

I. INTRODUCTION

The acoustic speech signal simultaneously conveys information regarding *who* is speaking and *what* is being said. Traditionally, these two functions were considered to be supported by different aspects of the acoustic signal, with indexical cues (e.g., fundamental frequency) used to support voice recognition and phonetic cues (e.g., voice-onset-time, formant patterns) used to support linguistic processing. We now know that a strict functional delineation between phonetic and acoustic cues is not possible. For example, talkers show stable individual differences in how they implement phonetic cues (e.g., Allen et al., 2003; Chodroff & Wilson, 2017; Hillenbrand et al., 1995; Theodore et al., 2009) and listeners are sensitive to these differences (e.g., Allen & Miller, 2004; Ganugapati & Theodore, 2019; Myers & Theodore, 2017; Theodore et al., 2015; Theodore & Miller, 2010). Experience with a talker's voice support advantages for linguistic processing (e.g., Nygaard et al., 1994; Nygaard & Pisoni, 1998), and experience with a given language supports advantages for voice processing (e.g., Goggin et al., 1991; Orena et al., 2015; Perrachione et al., 2011) – providing further evidence that the processing of phonetic and indexical aspects of the speech stream are intertwined.

Though behavioral evidence points to a tight integration between phonetic and indexical cues, the extant neuroimaging literature suggests disassociate hemispheric dominance for these two aspects of speech processing, with left temporal regions dominant for processing phonetic identity and right temporal regions dominant for processing voice identity (e.g., Formisano et al., 2008; Latinus et al., 2013; Bonte et al., 2014; Chang et al., 2010; Liebenthal et al., 2003; Myers, 2007; Belin & Zatorre, 2003; van Lancker et al., 1989). This may reflect different timescales for phonetic and indexical cues (e.g., Poeppel, 2003) and/or different functional tasks (e.g., von Kriegstein et al., 2003). Training studies have shown that perceptual learning of talker-specific

phonetic detail can alter hemispheric processing of phonetic cues. For example, Myers and Theodore (2017) exposed listeners to two talkers who differed in their characteristic VOTs. Following exposure, neural activation was measured using fMRI as listeners completed a phonetic identification task for VOT variants that were typical or atypical of each talker. Their results showed that right temporoparietal regions including the right middle temporal gyrus (rMTG) implicated in voice processing were sensitive to talker typicality. Moreover, a functional connectivity analysis showed greater connectivity between the rMTG and two regions in the left-hemisphere phonetic network (left postcentral gyrus, left middle temporal gyrus and left superior temporal sulcus) for talker typical compared to talker atypical VOT variants.

In addition, Francis and Driscoll (2006) reported evidence indicating that short-term perceptual training could induce a left ear advantage for using VOT as a cue to talker identification that results in faster talker identification decisions for stimuli presented to the left compared to the right ear, consistent with right hemisphere dominance for talker processing. The transmission of sound from the peripheral to central nervous system consists of contralateral auditory pathways; that is, auditory fibers that carry sound from the ear to the brain decussate such that monaural stimulation results in relatively strong activation in the contralateral hemisphere with relatively weaker activation of the ipsilateral hemisphere (e.g., Pickles, 1998; Jancke et al., 2002). This is not to say that sound detected by the left ear is only processed by the right hemisphere, but the contralateral nature of the auditory pathway has been successful exploited to measure hemispheric dominance using behavioral as opposed to neuroimaging methods (e.g., Kimura, 1967), as we describe further below. During training phase of the Francis and Driscoll (2006) study, listeners heard two sets of tokens, one with characteristically short VOTs (30 ms) and one with characteristically longer VOTs (50 ms). On each trial, a token was

presented binaurally and listeners were asked to identify which of two talkers produced that token. Feedback was provided to train listeners to associate the short VOTs as one talker and the longer VOTs as the other talker. Both sets of tokens were based on the speech of a single talker and thus indexical cues (e.g., fundamental frequency, vocal quality) were held constant between the two "talkers." Consequently, the two talkers *only* differed in their characteristic VOTs. During the pre- and post-test phases, listeners completed the same talker identification task with two key exceptions: (1) no feedback was provided and (2) stimuli were presented monaurally (i.e., either to the left or right ear on a given trial). Francis and Driscoll hypothesized that a left ear (i.e., right hemisphere) processing advantage would emerge at post-test for listeners who learned to process the phonetic property (i.e., VOT) as a cue to talker identification. Consistent with this hypothesis, reaction times to correct responses were on average 92 ms faster for stimuli presented to the left ear compared to the right ear at post-test. This left ear advantage was not present at pre-test, suggesting that it emerged as a consequence of learning during the training phase.

Though broadly consistent with the extant neuroimaging literature, the finding from Francis and Driscoll (2006) is striking in the context of the dichotic listening literature. Specifically, Francis and Driscoll observed a laterality effect (i.e., a left ear advantage) for stimuli that were presented *monaurally*. Hemispheric dominance for processing different types of acoustic signals has been measured through behavioral *dichotic* listening paradigms. In a traditional dichotic listening task, relative contributions of the left and right hemispheres are segregated by presenting a target stimulus to either the left or right ear in conjunction with a competing stimulus to the ear that does not receive the target stimulus (Kimura, 1967). Thus, during a dichotic listening task, there is simultaneous presentation of different stimuli to each

ear, with one ear receiving the target stimulus and the other ear receiving a competing stimulus. As reviewed in Hugdahl (2011), over 50 years of research using the dichotic listening paradigm has established both its utility as a behavioral method to measure brain laterality effects as well as a clear understanding of the importance of presenting dichotic stimuli – that is, a target and a competitor to opposite ears – in order to elicit laterality effects. Indeed, the latter point was established from the introduction of this paradigm (Kimura, 1967). Because Francis and Driscoll (2006) did not present a competing stimulus in the ear contralateral to the ear receiving the target stimulus, their task was not dichotic in nature, and thus it is perhaps surprising that the left ear processing advantage for talker identification was observed using a monaural listening task. Their finding may suggest that a competing stimulus in the contralateral ear of interest is extraneous for a task of this nature. Consistent with this interpretation, González et al. (2010) demonstrated that a left ear processing advantage for repetition-priming effects in a talker identification task was *strengthened* when noise was presented in the contralateral ear, but the competing stimulus was not necessary to induce the left ear advantage.

In addition, the finding from Francis and Driscoll (2006) bears revisiting due to several methodological and empirical points. First, the left ear advantage was observed in a very small sample of participants ($n = 8$). Small sample sizes alone are not a determinant of either research quality or reproducibility. Indeed, the "small-N" design, in which an extremely large number of observations are made on only a few participants, has a rich precedent in the psychophysics domain (Smith & Little, 2018). In some cases, small-N designs may even promote better power and inferential validity compared to large-N designs (Smith & Little, 2018). However, for traditional designs, such as that used in Francis and Driscoll (2006), small sample sizes can increase the likelihood of false positives in the literature just as they can decrease the ability to

detect true effects (e.g., Button et al., 2013). Second, these participants reflected those who met a learning criterion, defined as a minimum improvement in talker identification accuracy of 5% from pre- to post-test. While it is sensible to limit analyses to those who learned to associate VOT as a cue to talker identification given the nature of the hypothesis, no justification for the specific learning criterion was provided. Third, the statistical evidence for the key interaction between test phase and ear of presentation, while statistically significant, was only marginally so ($p = 0.04$). Fourth, the small sample ($n = 8$) who met the learning criterion reflected fewer than half of the total participants tested ($n = 18$). That is, most participants were *not* able to learn to associate VOT as a cue to talker identification, and this study did not reveal what factors may influence whether a given listener can learn to use phonetic properties to support talker identification.

For these reasons, the goal of the current work is two-fold. First, in each of two experiments, we conducted a high-powered replication of Francis and Driscoll (2006) in order to examine whether the left ear processing advantage for phonetic cues to talker identification would generalize to a larger sample. Second, all participants in experiment 1 completed four individual differences measures in addition to the primary talker identification task to identify potential predictors of talker identification performance. The talker identification task was modeled after the paradigm used in Francis and Driscoll (2006). In experiment 1, test stimuli were presented monaurally to either the left or right ear. In experiment 2, test stimuli were presented dichotically, with noise presented to the contralateral ear of the target stimulus. The four individual differences measures consisted of a flanker task, a pitch perception task, a category identification task, and a within-category discrimination task. We assessed inhibitory control (using a flanker task) and pitch perception given previous evidence linking both of these

158 constructs to talker identification ability (Theodore & Flanagan, 2020; Xie & Myers, 2015). For
159 example, increased inhibitory control has been positively associated with talker identification
160 accuracy (Theodore & Flanagan, 2020) and invoked as an explanatory mechanism for heighted
161 talker identification abilities in bilingual compared to monolingual children (Levi, 2018). On this
162 view, heighted talker identification may reflect a stronger ability to inhibit irrelevant information
163 (e.g., phonetic or other linguistic content) to instead focus on other aspects of the signal (e.g.,
164 fundamental frequency) for the purposes of talker identification. Pitch perception has also been
165 positively associated with talker identification and talker discrimination (Theodore & Flanagan,
166 2020; Xie & Myers, 2015). Using a flanker and pitch perception task to assess inhibitory control
167 and auditory acuity, respectively, supports the examination of individual differences in non-
168 speech abilities as potential predictors of performance in the current speech perception task. In
169 contrast, the category identification task assessed listeners' VOT voicing boundaries and
170 identification slopes (the latter as a measure of how categorically listeners perceived the voicing
171 contrast). The within-category discrimination task assessed listeners' perceptual acuity for VOT
172 specifically which is a logical precursor to learning to use VOT as a cue to talker identification.
173 If the left ear advantage for phonetic cues to talker identification generalizes beyond the
174 original sample (Francis & Driscoll, 2006), then we predict that listeners who learn to associate
175 VOT as a cue to talker identification will show faster reaction times for stimuli presented to the
176 left ear compared to the right ear during the talker identification post-test. If increased inhibitory
177 control and auditory acuity are associated with enhanced talker identification, then we predict a
178 positive relationship between performance on the talker identification task and performance on
179 the flanker, pitch perception, and within-category discrimination tasks. The relationship between
180 talker identification and categorical perception was exploratory in the current work. Listeners

who show early VOT voicing boundaries may not have perceived the VOT variants in the talker identification task as belonging to the same category, which would result in improved talker identification accuracy given that the two talkers would be perceived as saying different words (instead of saying the same word with different VOTs). Listeners who have shallower identification slopes may be more sensitive to within-category variation compared to listeners who have more categorical slopes; if so, then those with shallower identification slopes would show improved performance on the talker identification task.

II. EXPERIMENT 1

A. Methods

1. Participants

One hundred and forty participants were recruited from the Prolific participant pool (https://www.prolific.co; Palan & Schitter, 2018) for session one. All participants were monolingual English speakers between 18 – 35 years of age currently residing in the US with no history of language-related disorders. Forty-three participants were excluded due to failure to pass all three headphone screens (n = 28) or failure to meet the training accuracy criterion (n = 15), described in detail below. The final sample (n = 97) included 42 women and 55 men (*mean* age = 27 years, SD = 4 years). All of these participants were invited to participate in session two, with 59 participants choosing to do so. The mean time between the two sessions was 11 days (SD = 12 days, *range* = 1 – 35 days).

2. Power analysis

The sample size was determined based on a priori power analyses using the simr package (Green & MacLeod, 2016) in R. First, trial-level data from Francis and Driscoll (2006) for the effect we aimed to replicate (i.e., the data underlying the interaction shown in their Figure 1)

were fit to a linear mixed effects model using the lmer() function from the lme4 package (Bates et al., 2015) in R. The dependent variable was log-transformed reaction time. The fixed effects were test (pre-test = -0.5, post-test = 0.5), ear (left ear = -0.5, right ear = 0.5), and their interaction. The random effects structure consisted of random intercepts by subject and random slopes by subject for test and ear. Second, we created a data frame to reflect the structure of our design, given that power for mixed effects model is linked to number of observations (in addition to sample size). As described below, each participant completed 80 trials in each test phase, and we only analyzed RTs for correct responses (as in Francis and Driscoll, 2006). We conservatively simulated accuracy at 60% correct, resulting in a simulated data set that consisted of 48 trials/participant at each test session. Third, the parameters of the original model were simulated in our data structure 500 times using the powerCurve() function in simr, which showed that 55 participants were required to achieve 80% power to detect the test by ear interaction observed in Francis and Driscoll (2006). Thus, we aimed to test 55 participants who met the learning criterion from Francis and Driscoll (2006) in order to perform an adequately powered replication. The recruited sample (n = 140) was based on estimated attrition rates (e.g., failure to pass headphone screens, failure to pass training criterion, failures to meet learning criterion) and resulted in 58 participants who met the learning criterion from Francis and Driscoll (2006), as we describe further below.

3. Stimuli

a. Talker identification. Stimuli for the talker identification task were drawn from two VOT continua, one that perceptually ranged from *gain* to *cane* and one that perceptually ranged from *goal* to *coal*. Both continua were created by applying an LPC synthesis procedure to natural productions of the voiced endpoints (i.e., *gain*, *goal*) elicited from a single female, monolingual

speaker of American English. These stimuli are a subset of those used in Theodore and Miller (2010), to which the reader is referred to for comprehensive details on stimulus construction.

From each of these continua, we selected two unique tokens for each of the two talkers, who were fictitiously referred to as Joanne and Sheila. The selected tokens were in the unambiguous, voiceless region of the original continua and thus perceptually cued the words *cane* and *coal*. The tokens were selected so that Joanne had characteristically shorter VOTs than Sheila. Specifically, the selected VOTs ranged between 84 – 89 ms for Joanne and between 165 – 170 ms for Sheila. With this procedure, the only difference between the two talkers' voices was their characteristic VOTs. These stimuli were used for both the training and test phases because the left ear advantage in Francis and Driscoll (2006) only emerged for trained items. As described in the procedure section below, stimuli were presented binaurally during training and monaurally at test.

b. Individual differences measures. Separate stimulus sets were used in each of the four individual differences tasks. Stimuli for the flanker task consisted of linear arrays of five arrows in which the middle arrow was either congruent (e.g., < < < < <) or incongruent (e.g., < < < > < <) with the flanking arrows. There were 80 arrays in total, 20 congruent and 20 incongruent arrows for each of two arrow directions (i.e., left vs. right). Stimuli for the pitch perception task consisted of a subset of the local pitch perception stimuli from Xie and Myers (2015), to which the reader is referred for comprehensive details on stimulus construction. In brief, each stimulus consisted of a pair of six-tone sequences separated by 1000 ms of silence. There were 32 pairs in total; 16 of which contained two identical tone sequences (i.e., same trials) and 16 of which contained tone sequences that differed in pitch for one of the six tones of the sequence (i.e., different trials).

Stimuli for the category identification and within-category discrimination tasks consisted of tokens drawn from *gain-cane* and *goal-coal* VOT continua, created using the methods described for the talker identification stimuli. Critically, the stimuli used for the category identification and within-category discrimination tasks were produced by a different talker than was used in the main talker identification learning task in order to minimize potential transfer of learning between the two sessions. Stimuli for the category identification task consisted of 10 tokens from each continuum consisting of VOTs that ranged between 21 – 99 ms; as a consequence, the selected VOTs perceptually cued both endpoints for each continuum (i.e., *gain*, *cane*, *goal*, *coal*). Stimuli for the within-category identification task consisted of 15 tokens from each continuum consisting of VOTs that ranged between 79 – 208 ms; accordingly, all selected tokens cued the voiceless endpoint (i.e., *cane*, *coal*). The selected tokens were arranged into same and different pairs; word was held constant on a given pair. There were 12 unique same pairs that sampled the range of selected VOTs. There were 36 different pairs, reflecting six unique pairs for each of three step distances between pair members (reflecting a difference in VOT of 28, 54, or 80 ms, respectively) and two pair orders. As we describe in the procedure section, we presented three repetitions of the same pairs and two repetitions of the different pairs during the within-category discrimination task in order to equate the number of same and different trials.

4. Procedure

a. Session 1. All testing was completed online using Gorilla Experiment Builder https://gorilla.sc; Anwyl-Irvine et al., 2019). After providing informed consent, participants completed a series of headphone screens. These included two existing protocols that use dichotic listening tasks to screen for headphone compliance on web-based platforms (Milne et al., 2021;

Woods et al., 2017). The third screen was a custom channel detection task in which listeners heard a tone presented to either the left or right ear and were asked to indicate via button press in which ear they heard the tone. The channel detection task was used in the headphone screen battery because although the Woods et al. (2017) and Milne et al. (2020) dichotic listening tasks assess use of stereo headphones, they do not assess whether a participant has placed the left channel on the left ear (and thus the right channel on the right ear), which is a critical requirement for the present study. Participants who did not pass on all three screens were excluded from analyses; pass was defined as ≥ 5 correct responses (of six total trials) on the Woods et al. (2017) and Milne et al. (2020) tasks and ≥ 7 correct responses (of eight total trials) on the custom channel detection task.

After completing the headphone screens, participants completed the talker identification task. The talker identification task consisted of familiarization, pre-test, training, and post-test phases. During familiarization, listeners heard two repetitions of the four tokens for each talker while seeing the name of the talker displayed on the screen. Stimuli during familiarization were presented binaurally and blocked by word and talker (i.e., they heard Joanne's two *cane* tokens, and then Sheila's two *cane* tokens, followed by Joanne's two *coal* tokens and then Sheila's two *coal* tokens). Listeners were directed to listen to each word, view the talker's name, and try to learn each talker's voice; no responses were collected during familiarization.

Following familiarization, listeners completed the pre-test phase. On a given trial, stimuli were presented monaurally to either the left or right channel. The pre-test consisted of 80 trials (2 tokens x 2 words x 2 talkers x 2 channels x 5 repetitions) presented in a different randomized order for each participant. Participants were asked to indicate whether the word was produced by Joanne or by Sheila as quickly as possible without sacrificing accuracy. Participants made their

responses using the “a” and “l” keys and were instructed to keep their index fingers on these keys throughout the experiment to facilitate faster response times; a visual diagram was provided to demonstrate correct finger placement during the instructions. The instructions explicitly noted that this was not a sound localization task to help ensure that participants understood that they should be indicating which talker they heard on each trial and not ear of stimulus presentation. No feedback was provided during pre-test.

After the pre-test, participants completed the training phase. The training phase consisted of 400 trials (2 tokens x 2 words x 2 talkers x 50 repetitions) of talker identification following the task instructions described for pre-test; trials were randomized separately for each participant. Stimuli were presented binaurally and feedback was provided on every trial in the form of a green checkmark (for correct responses) or a red “x” (for incorrect responses). Session one concluded with the post-test phase, which was identical to the pre-test phase.

A progress bar was displayed on the bottom center of the screen throughout the entirety of the experiment and the ISI was constant at 1000 ms (measured from the participant’s response on each trial to the onset of the next stimulus). The entire procedure lasted approximately 35 minutes and participant were compensated $5.83 for their participation.

b. Session 2. All testing was completed online using Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). After providing informed consent, listeners completed the Woods et al. (2018) and Milne et al. (2020) headphone screens. All participants who returned for session two (n = 59) passed all headphone screens at session two. After completing the headphone screens, participants completed the within-category discrimination, flanker, category identification, and pitch perception tasks in this fixed order. The within-category discrimination consisted of 72 same trials (2 words x 12 VOTs x 3 repetitions) and 72 different trials (2 words x 3 distances x 6

unique pairs x 2 pair orders). On each trial, participants were directed to indicate whether the two members of the pair were the same or different by clicking one of two appropriately labeled buttons. The flanker task consisted of one randomization of the 80 linear arrays (2 trial types x 2 directions x 20 repetitions). On each trial, participants were directed to indicate the direction of the central arrow as quickly as possible without sacrificing accuracy. Participants were asked to keep their index fingers on top of the response keys throughout the task and a visual diagram illustrating correct finger placement was provided during the instructions.

The category identification task consisted of 80 trials (2 continua x 10 VOTs x 4 repetitions). Participants were asked to indicate whether the word began with a "g" as in *gain* and *goal* or "k" as in *cane* and *coal* by clicking on an appropriately labeled button. The pitch perception task consisted of 64 trials (2 trial types x 16 unique pairs x 2 repetitions). On each trial, participants indicated whether the two members of the pair were the same or different by clicking on an appropriately labeled button. For all tasks, trials were presented in a separate randomized order for each participant and the ISI was 1000 ms. The entire procedure lasted approximately 35 minutes; participants were compensated $5.83 for their participation.

B. Results

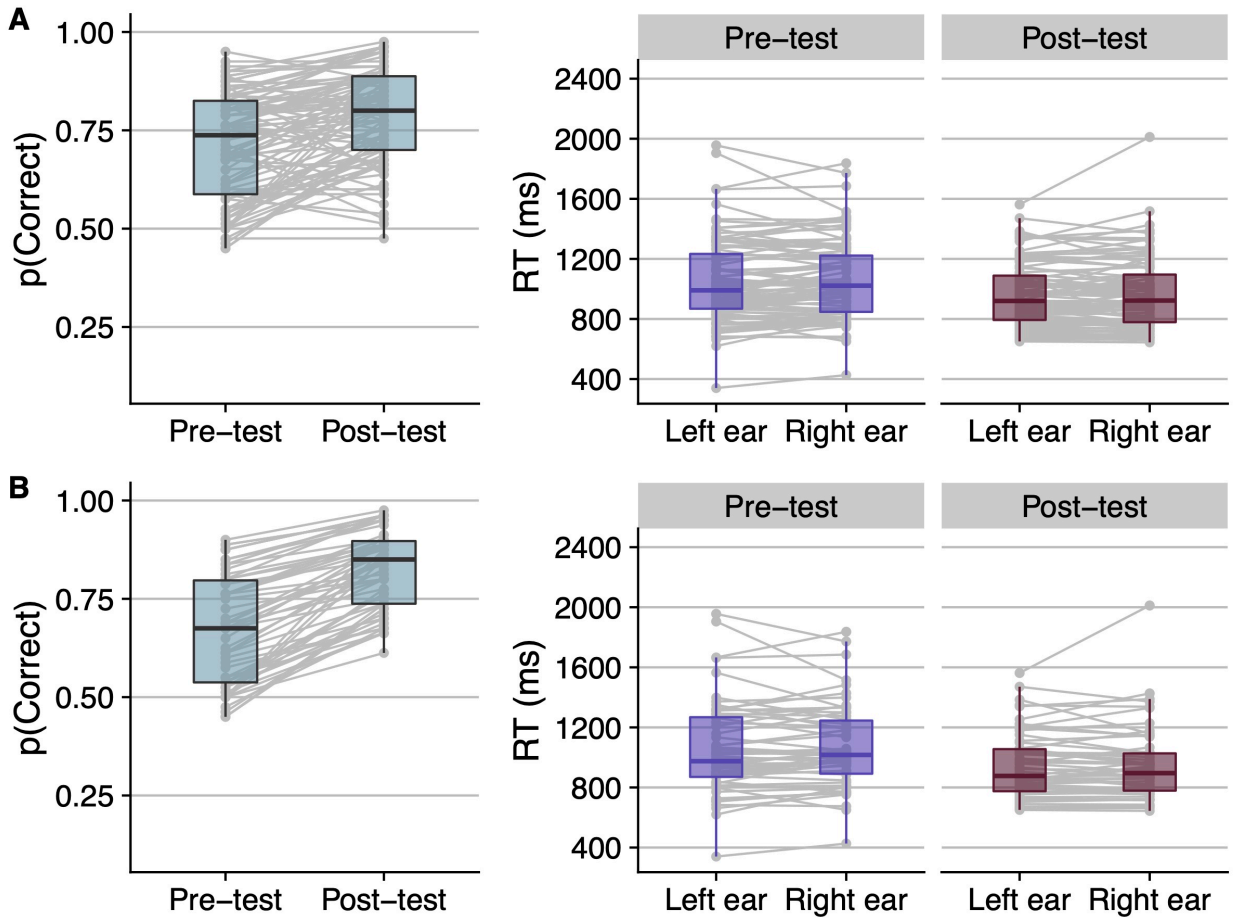
1. Talker identification

a. Training. Trial-level data (for all tasks) and a script (in R) is available on the Open Science Framework (https://osf.io/ge2vb/). Executing the script will reproduce all statistics reported in this manuscript in addition to generating all figures. For the training phase, accuracy for each participant was calculated in terms of proportion correct responses across all training trials. We excluded 15 participants because they failed to meet the inclusion criterion for training accuracy (≥ 0.60). Mean accuracy across included participants (0.83, $SD = 0.10$, *range* = 0.61 –

342 0.98) was significantly above chance as confirmed by a one-sample t-test [$t(96) = 32.495$, $p <$
343 0.001), which was expected based the inclusion criterion.

344 *b. Test.* Accuracy and reaction time during the test phases were analyzed separately. For
345 accuracy, trial-level responses (0 = incorrect, 1 = correct) were submitted to a generalized linear
346 mixed effects model with the binomial response family as implemented with the glmer() function
347 of the lme4 package (Bates et al., 2015) in R. The model included fixed effects of test (pre-test =
348 -0.5, post-test = 0.5), ear (left = -0.5, right = 0.5), and their interaction. The random effects
349 structure included random intercepts by subject and random slopes by subject for test and ear.
350 The model revealed a main effect of test ($\hat{\beta} = 0.469$, $SE = 0.069$, $z = 6.829$, $p < 0.001$), indicating
351 that accuracy improved from pre-test (0.71, $SD = 0.14$) to post-test (0.79, $SD = 0.12$). There was
352 no main effect of ear ($\hat{\beta} = 0.034$, $SE = 0.043$, $z = 0.780$, $p = 0.435$), nor an interaction between
353 test and ear ($\hat{\beta} = -0.090$, $SE = 0.078$, $z = -1.150$, $p = 0.250$). The main effect of test is visualized
354 in Figure 1.

355 **FIG. 1.** (Color online.) Results of the talker identification task in experiment 1. Panel A shows
356 performance during the talker identification task for all participants ($n = 97$); panel B shows
357 performance during the talker identification for those who met the learning criterion ($n = 58$). In
358 both panels, the distribution of participants' accuracy scores (mean proportion correct) for each
359 test is shown at left, and the distribution of participants' mean response times to correct
360 responses by test and ear of stimulus presentation is shown at right.



Trial-level reaction times (RT) for correct responses during test were analyzed in a linear mixed effects model using the lmer() function of the lme4 package (Bates et al., 2015). The Satterthwaite approximation of degrees of freedom was used to evaluate statistical significance using the *t* distribution (Kuznetsova et al., 2017). RTs were log-transformed and trials exceeding 2.5 SDs of a participant's mean log RT were excluded (3.2% of correct RTs). The fixed and random effects structure was identical to that described for the accuracy analysis. The results of the model showed a main effect of test ($\hat{\beta}$ = -0.073, *SE* = 0.020, *t* = -3.620, $p < 0.001$), with RTs decreasing from pre-test (*mean* = 1051 ms, *SD* = 268) to post-test (*mean* = 958 ms, *SD* = 220). There was no main effect of ear ($\hat{\beta}$ = 0.008, *SE* = 0.005, *t* = 1.489, $p = 0.140$), nor an interaction between test and ear ($\hat{\beta}$ = -0.011, *SE* = 0.010, *t* = -1.075, $p = 0.283$). Figure 1 shows the

distribution of participants' mean RTs by test session and ear.

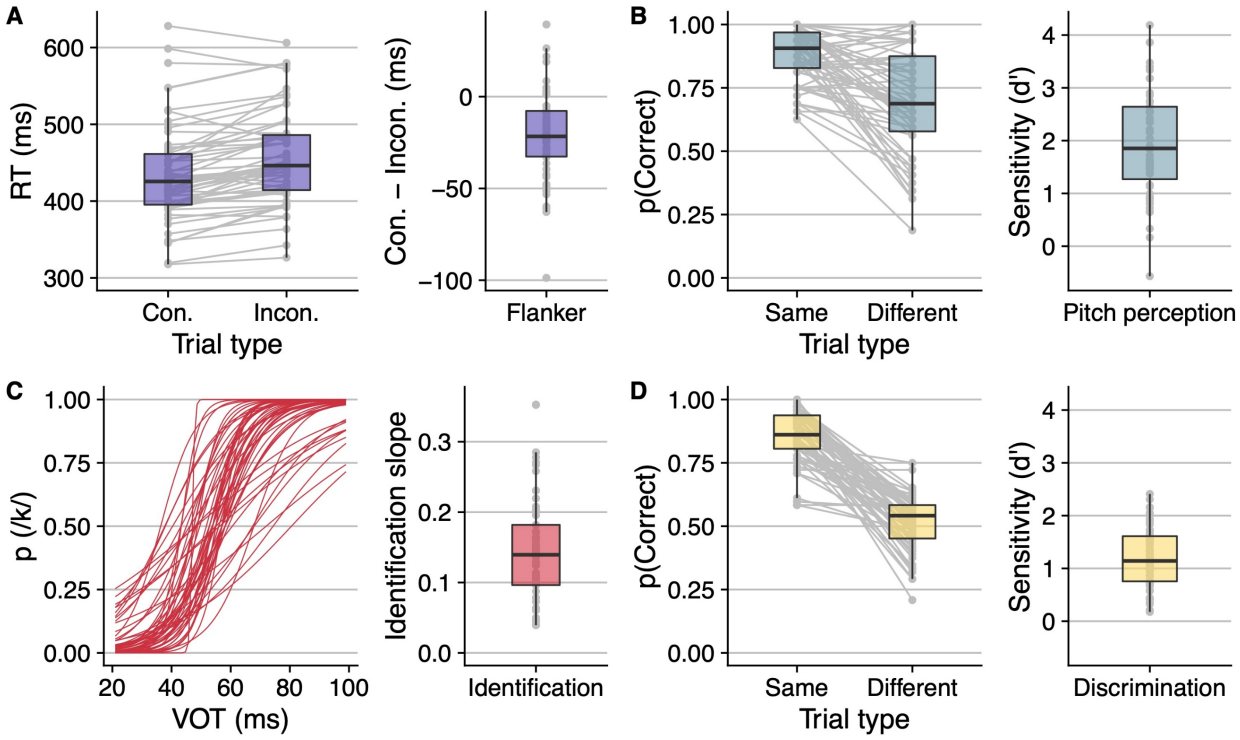
Recall that in Francis and Driscoll (2006), the left ear advantage at post-test emerged only for listeners who met their learning criterion, defined as ≥ 5% improvement in talker identification accuracy from pre-test to post-test. A parallel reaction time analysis was performed limited to listeners in the current study who met this criterion (n = 58). The results converged with the full sample; reaction time decreased from pre-test to post-test ($\hat{\beta}$ = -0.083, SE = 0.031, t = -2.671, p = 0.010), but there was no main effect of ear ($\hat{\beta}$ = 0.008, SE = 0.007, t = 1.152, p = 0.255), nor an interaction between test and ear ($\hat{\beta}$ = -0.020, SE = 0.013, t = -1.541, p = 0.123).

2. Individual differences measures

a. Flanker. Mean accuracy (proportion correct) across participants was near ceiling (0.97, SD = 0.03, *range* = 0.85 – 1.00). To ensure that the expected inhibition effect was observed across participants in the aggregate, trial-level log RTs for correct responses were analyzed in a linear mixed effects model following the methods outlined previously. RTs were log-transformed and trials exceeding 2.5 SDs of a participant's mean log RT were excluded (2.8% of correct RTs). Congruency was entered as a fixed effect (congruent = -0.5, incongruent = 0.5); the random effects structure consisted of random intercepts by participant and random slopes for congruency by participant. The results of the model confirmed a main effect of congruency ($\hat{\beta}$ = 0.049, SE = 0.006, t = 7.648, p < 0.001), with RTs faster for congruent (*mean* = 434 ms, SD = 63) compared to incongruent trials (*mean* = 455 ms, SD = 61). For each subject, inhibition was calculated as the difference in RT between congruent and incongruent trials; thus, more negative scores indicate weaker inhibition. Performance on the four individual differences measures is shown in Figure 2.

FIG. 2. (Color online.) Performance on the four individual differences measures in experiment 1. Panel A shows the distribution of participants' mean response times by trial type (left) and the

distribution of interference scores (right) for the flanker task. Panel B shows the distribution of accuracy scores for same and different trials (left) and the distribution of sensitivity scores (right) for the pitch perception task. Panel C shows the relationship between /k/ responses and VOT for each participant as determined by logistic regression (left) and the distribution of identification slopes (right) for the category identification task. Panel D shows the distribution of accuracy scores for same and different trials (left) and the distribution of sensitivity scores (right) for the within-category discrimination task.



b. Pitch perception. To quantify performance on the pitch perception task, we calculated sensitivity (*d'*) separately for each participant. Hit was defined as responding "same" for same tone sequence trials; false alarm was defined as responding "same" for different tone sequence trials. Mean sensitivity (*d'*) across participants was 1.96 (*SD* = 1.01), which was significantly greater than zero [$t(58) = 14.847$, $p < 0.001$].

c. Category identification. Trial-level identification responses were fit to a logistic regression separately for each participant; in each regression, VOT was the independent variable and binary response (0 = /g/, 1 = /k/) was the dependent variable. Two parameters were derived from each regression: (1) the slope of the identification function and (2) the category boundary,

defined as the VOT corresponding to 0.50 proportion /k/ responses. To derive the slope, we used the beta estimate for VOT from the regression model; with this metric, higher values indicate steeper identification slopes. The category boundary was derived using the model intercept and beta estimate for VOT from the regression model according to equation (1), where $\hat{\beta}_0$ is the intercept, $\hat{\beta}_1$ is the slope, and x is the category boundary.

$$\hat{\beta}_0 + \hat{\beta}_1 x = \log(\frac{0.5}{1-0.5}); \; x = -\frac{\hat{\beta}_0}{\hat{\beta}_1} \qquad \text{Equation (1)}$$

Two participants were excluded from subsequent category identification analyses because they did not show a statistically significant relationship between VOT and phonetic decisions; instead, their response functions were flat suggesting that they did not perform the task as directed. Across participants, the mean slope of the identification function was 0.146 (*SD* = 0.076) and the mean category identification boundary was 54 ms (*SD* = 9 ms).

d. Within-category discrimination. To quantify performance on the within-category discrimination task, we calculated sensitivity (*d'*) separately for each participant. Hit was defined as responding "same" for same trials; false alarm was defined as responding "same" for different trials. Mean sensitivity (*d'*) across participants was 1.20 (*SD* = 0.56), which was significantly greater than zero [$t(58) = 16.348$, $p < 0.001$].

3. Relationship between talker identification and individual differences measures

A series of correlations were performed to examine whether performance on the individual differences measures predicted performance on the talker identification task. Five measures of individual differences were considered, which included inhibition (congruent RT – incongruent RT), sensitivity (d') for the pitch perception task, identification slope and category boundary for the category identification task, and sensitivity (d') for the within-category discrimination task. Each of these five measures was correlated with four measures from the

434 talker identification task, which included accuracy during training, accuracy at pre-test, accuracy
435 at post-test, and learning (accuracy at post-test – accuracy at pre-test). These correlations are
436 shown in Table I and visualized in Figure 3.

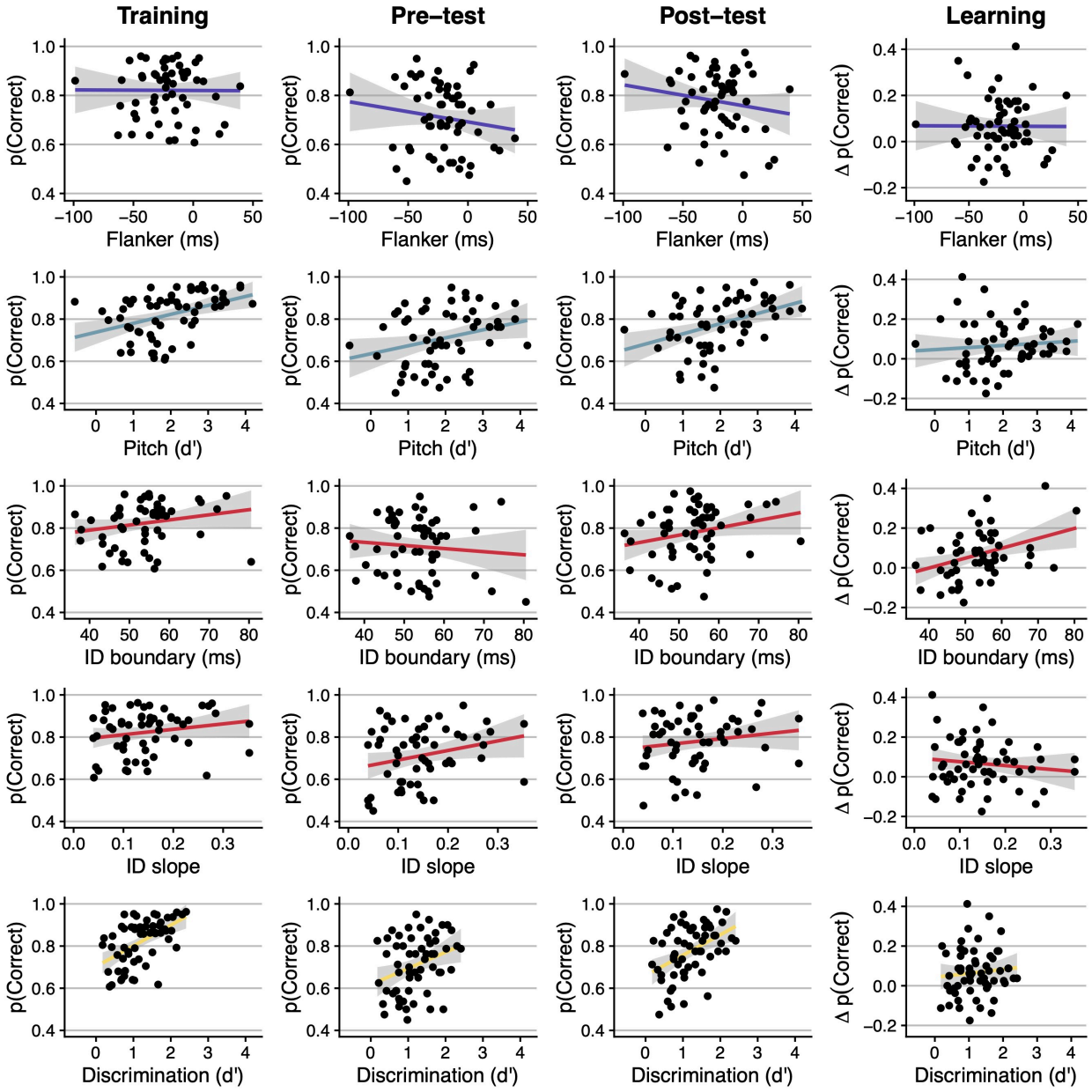
437 **TABLE I**. Pearson's correlation coefficient (*r*) and p-value (in parentheses) relating the five
438 individual difference measures to each of four talker identification measures. Cells in bold
439 indicate statistical significance with $\alpha = 0.05$. Cells with underline indicate statistical
440 significance after applying the conservative Bonferroni correction to account for family-wise
441 error rate (as described in the main text). The degrees of freedom were 57 for all measures except
442 for the category identification measures, which had degrees of freedom equal to 55 (given that
443 two participants were excluded due to failure to complete the task as directed). Parallel analyses
444 were performed using Spearman's rank-order correlations and the results converged in all cases;
445 these correlations can be viewed by executing the script provided in the OSF repository for this
446 manuscript.

Individual differences measure	Talker identification measure			
	Training	Pre-test	Post-test	Learning
Inhibition	0.00 (0.971)	-0.15 (0.267)	-0.17 (0.209)	0.00 (0.971)
Pitch perception	**0.41 (0.001)**	**0.28 (0.029)**	**0.40 (0.002)**	0.09 (0.500)
Identification: Boundary	0.20 (0.132)	-0.09 (0.487)	0.25 (0.065)	**0.36 (0.007)**
Identification: Slope	0.19 (0.165)	0.25 (0.058)	0.16 (0.240)	-0.12 (0.363)
Discrimination	**0.52 (<0.001)**	**0.32 (0.013)**	**0.45 (<0.001)**	0.09 (0.477)

447 There was no significant relationship between inhibition and any measure of talker
448 identification. Pitch perception was positively associated with talker identification accuracy
449 during training, pre-test, and post-test; however, pitch perception was not related to the degree of
450 learning. The location of the VOT voicing boundary was significantly associated with the degree
451 of learning from pre- to post-test; with longer category boundaries associated with better
452 performance. Within-category discrimination was positively associated with talker identification
453 accuracy during training, pre-test, and post-test, but was not associated with learning.

454 **FIG. 3**. (Color online.) Scatterplots illustrating the relationship between the five individual
455 differences measures (by row) and the four measures of talker identification (by column) in
456 experiment 1. Each point reflects an individual participant. The regression line indicates a linear

model; the shaded region marks the 95% confidence interval.



We note that when the conservative Bonferroni correction to account for family-wise error rate is applied (resulting in corrected $\alpha = 0.0025$ given $\alpha = 0.05$ and 20 comparisons), the only relationships that survive are the associations between the two measures of auditory acuity

461 (i.e., pitch perception, within-category discrimination) and talker identification accuracy during
462 training and post-test.[1]

463 **III. EXPERIMENT 2**

464 The results of experiment 1 revealed two primary findings. First, most listeners learned to
465 use VOT as a cue to talker identification, consistent with the results of Francis and Driscoll
466 (2006). That is, following a brief training phase, listeners improved in their ability to use a
467 phonetic cue as an indicant of talker identity, even in the absence of traditional indexical cues to
468 voice identity (e.g., fundamental frequency). Second, auditory acuity was positively associated
469 with talker identification, suggesting that heightened sensitivity to fine-grained acoustic
470 information facilitated performance in the current task. Of note, we did not observe any evidence
471 to suggest that ear of stimulus presentation influenced performance in the talker identification
472 task. The goal of experiment 2 is examine whether a left ear advantage is observed under
473 conditions that are known to better facilitate behavioral observation of laterality effects.
474 Following the conclusion of experiment 2, we present Bayes Factors analyses to inform
475 interpretation of null effects reported in this manuscript.

476 As reviewed in the introduction, hemispheric laterality effects are more optimally
477 observed in behavioral tasks under conditions in which a competing stimulus is presented to the
478 contralateral ear of the target stimulus (Behne et al., 2005, 2006; Bless et al., 2015; González et
479 al., 2010; Hugdahl & Anderson, 1984; Studdert-Kennedy & Shankweiler, 1970; Westerhausen,

[1] In addition to packages cited in the main text, we also acknowledge additional R resources used for data analysis including the tidyverse packages dplyr and tidyr (Wickham et al., 2019) for data manipulation, the tidyverse package ggplot2 (Wickham et al., 2019) and the cowplot package (Wilke, 2019) for figure generation, the jtools package (Long, 2020) for summarizing model results, and the inauguration package (Bedford-Petersen, 2021), which provides a color palette for plots inspired by the power attire worn by celebrated women at the 2021 US presidential inauguration.

2019). This is contrast to the manipulation used in Francis and Driscoll (2006) and in experiment 1, in which silence was presented to the contralateral ear. Dichotic stimulus presentation facilitates the observation of laterality effects because ipsilateral auditory pathways are suppressed when ears are presented with competing stimuli. Most of the literature on laterality effects for auditory verbal processing differs from the focus of the current investigation in that previous research has primarily examined laterality effects when processing the linguistic content of the stimuli, that is, the "what" of a talker's message. For example, the pioneering work of Kimura (1967) presented verbal productions of different digits to each ear and asked listeners to identify which digit(s) they heard. Likewise, the now classic consonant-vowel (CV) dichotic listening paradigm presents different CV syllables to each ear and requires listeners to identify which syllables(s) they hear (Hugdahl & Anderson, 1984; Studdert-Kennedy & Shankweiler, 1970). The extensive literature on *linguistic* processing of dichotic signals thus supports a cumulative science that can inform optimal design decisions for eliciting and measuring laterality effects for auditory verbal processing (e.g., Bless et al., 2013, 2015; Parker et al., 2021; Westerhausen, 2019).

In contrast, studies using behavioral tasks to assess hemispheric laterality for talker processing – that is, the "who" of a linguistic message – is relatively sparse, reflecting an emerging line of inquiry. To our knowledge, only three studies have provided behavioral evidence of a left ear advantage for talker identification. The first is Francis and Driscoll (2006), which directly motivates the current work, and, as described previously, consisted of a monaural task (i.e., stimuli presented to either the left or right ear) instead of a dichotic listening task. The second comes from Perrachione and colleagues (Perrachione et al., 2009). In their study, native English and native Mandarin listeners completed a talker identification task (with feedback) for

voices speaking English and Mandarin during a training phase. On each trial, listeners heard two talkers produce the same sentence and were asked to identify the talker in the left ear on some trials and the talker in the right ear on other trials. Trials were blocked by both ear and stimulus language; that is, listeners completed four training blocks formed by crossing monitoring ear (left vs. right) and stimulus language (English vs. Mandarin). Analysis of talker identification accuracy during training revealed a left ear benefit for both listener groups only when identifying talkers producing English sentences, which the authors speculate may reflect differences in the temporal modulation of frequency information between the two languages.

The third study comes from González and colleagues (González et al., 2010). In their study, listeners completed a talker identification task with target stimuli presented to either the left or the right ear. The construct of interest in this study was long-term repetition priming; accordingly, talker identification accuracy was compared between same sentence (i.e., a talker's repeated sentence) and different sentence (i.e., a talker's novel sentence) trials. Pink noise was presented in the contralateral ear to the target stimulus in their first experiment, whereas silence was presented in the contralateral ear in their second experiment. The results of the two experiments converged to show a left ear advantage for recognition memory in the talker identification task. Specifically, talker identification accuracy was higher for same compared to different sentence trials when stimuli were presented in the left ear, and no such benefit was observed for stimuli presented in the right ear. The laterality effect was observed in both experiments; however, it was stronger in the first compared to the second experiment, consistent with noise in the contralateral ear serving to suppress the influence of ipsilateral auditory pathways (Behne et al., 2005, 2006).

Drawing from these three studies, the specific dichotic manipulation used in experiment 2

was to present pink noise in the contralateral ear to the target stimulus, as in González et al. (2010). This manipulation allowed us to use otherwise identical procedures between experiments 1 and 2, and thus better isolate the influence of a dichotic listening environment on any observed differences between the two experiments (in contrast to, for example, adopting the blocked ear design used in Perrachione et al., 2009).

A. Methods

1. Participants

One hundred and fourteen participants were recruited from the Prolific participant pool (https://www.prolific.co; Palan & Schitter, 2018) following the criteria outlined for experiment 1. Forty-three participants were excluded due to failure to pass all three headphone screens (n = 23) or failure to meet the training accuracy criterion (n = 12), as described for experiment 1. The final sample (n = 79) included 24 women, 54 men, and one participant who declined to report gender (*mean* age = 28 years, *SD* = 4 years). The sample size was determined by the power analyses described for experiment 1. Specifically, we tested participants until we achieved n = 55 who met the learning criterion outlined for experiment 1, which was defined as an improvement in proportion correct talker identification between pre- and post-test greater than or equal to 0.05.

2. Stimuli and procedure

The stimuli and procedure were identical to experiment 1 with two key exceptions. First, pink noise was presented in the contralateral ear of the target stimulus at pre- and post-test. Following González et al. (2010), the amplitude of the pink noise (72 dB) was 3 dB lower than the amplitude of the target stimuli (75 dB). Second, in addition to the 80 target trials in each test phase (i.e., trials on which the target was presented to either the left or the right ear, with pink noise presented in the contralateral ear), 20 filler trials were presented in which the target

549 stimulus was presented binaurally (i.e., the same signal was presented to each ear), following
550 recommendations of Parker et al. (2021) and Westerhausen (2019). These filler trials, reflecting
551 five repetitions of each word for each talker, were randomly interspersed across each test phase
552 and subsequently removed from the analyses.

553 **B. Results**

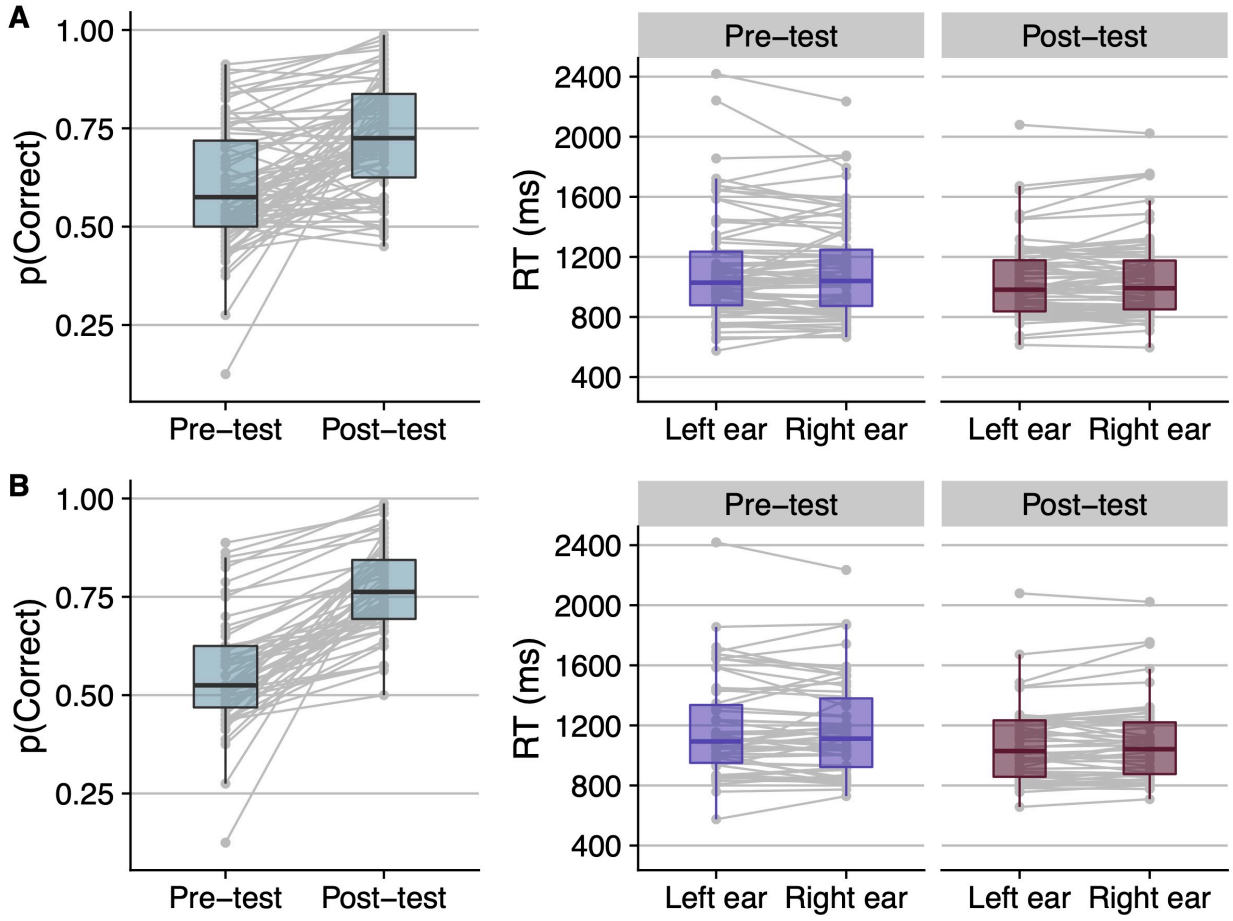
554 Performance during the training phase was analyzed as outlined for experiment 1. Mean
555 accuracy across participants (0.82, *SD* = 0.10, *range* = 0.62 – 0.98) was significantly above
556 chance as confirmed by a one-sample t-test [$t(78) = 27.492$, $p < 0.001$), which was expected
557 based the inclusion criterion (accuracy ≥ 0.60).

558 Accuracy and reaction time during the test phases were analyzed as outlined for
559 experiment 1. The accuracy model revealed a main effect of test ($\hat{\beta} = 0.661$, $SE = 0.095$, $z =$
560 6.942, $p < 0.001$), indicating that accuracy improved from pre-test (0.60, *SD* = 0.15) to post-test
561 (0.73, *SD* = 0.14). There was no main effect of ear ($\hat{\beta} = -0.041$, $SE = 0.042$, $z = -0.985$, $p =$
562 0.325), nor an interaction between test and ear ($\hat{\beta} = 0.028$, $SE = 0.080$, $z = -0.340$, $p = 0.734$).
563 The main effect of test is visualized in Figure 4.

564 As described for experiment 1, RTs were log-transformed and trials exceeding 2.5 SDs of
565 a participant's mean log RT were excluded from analysis (2.7% of correct trials). The results of
566 the RT model showed a main effect of test ($\hat{\beta} = -0.048$, $SE = 0.022$, $t = -2.175$, $p = 0.033$), with
567 RTs decreasing from pre-test (*mean* = 1112 ms, *SD* = 330) to post-test (*mean* = 1041 ms, *SD* =
568 257). There was no main effect of ear ($\hat{\beta} = 0.011$, $SE = 0.006$, $t = 1.716$, $p = 0.090$), nor an
569 interaction between test and ear ($\hat{\beta} = 0.006$, $SE = 0.012$, $t = 0.476$, $p = 0.634$). Figure 4 shows the
570 distribution of participants' mean RTs by test session and ear.

571 **FIG. 4.** (Color online.) Results of the talker identification task in experiment 2. Panel A shows
572 performance during the talker identification task for all participants ($n = 79$); panel B shows

performance during the talker identification for those who met the learning criterion (n = 55). In both panels, the distribution of participants' accuracy scores (mean proportion correct) for each test is shown at left, and the distribution of participants' mean response times to correct responses by test and ear of stimulus presentation is shown at right.



Recall that in Francis and Driscoll (2006), the left ear advantage at post-test emerged only for listeners who met their learning criterion, defined as ≥ 5% improvement in talker identification accuracy from pre-test to post-test. As for experiment 1, a parallel reaction time analysis was performed limited to listeners in experiment 2 who met this criterion (n = 55). The results converged with the full sample; reaction time decreased from pre-test to post-test ($\hat{\beta}$ = -0.078, SE = 0.027, t = -2.859, p = 0.006), but there was no main effect of ear ($\hat{\beta}$ = 0.009, SE = 0.007, t = 1.213, p = 0.225), nor an interaction between test and ear ($\hat{\beta}$ = 0.003, SE = 0.015, t = 0.235, p = 0.815).

IV. BAYES FACTORS ANALYSIS

Collectively, the results of experiment 2 converge with the pattern of results observed for the talker identification task in experiment 1. Specifically, most listeners learned to use a phonetic property of speech as a cue to talker identification, as indicated by an improvement in talker identification accuracy from pre- to post-test. In addition to improved accuracy, exposure during the training phase inferred a behavioral benefit such that talker identification decisions were faster at post-test compared to pre-test. However, there was no evidence to suggest that learning to use a phonetic property as a cue to talker identity yielded a left ear advantage for talker identification, even under circumstances that were optimized to elicit a laterality effect in behavior (i.e., by presenting a competing stimulus in the contralateral ear to the target stimulus).

Drawing conclusions from a null result using frequentist statistics (i.e., null hypothesis significance testing) is challenging because, by definition, a *p*-value does not provide evidence in support of the null hypothesis. Instead, the *p*-value obtained in a frequentist analysis approach, as used in the current work, reflects the probability of observing the result (or a more extreme result) if the null hypothesis were true (e.g., Badenes-Ribera et al., 2016; Hubbard & Lindsay, 2008). That is, the *p*-value reflects the probability of the data given the null, which can be formally expressed as $p = p(data|H0)$. When the *p*-value is low (e.g., $p < 0.05$), we inferentially reason that we can reject the null hypothesis because the probability of the observed effect in the data is very low if in fact the null hypothesis were true. When the *p*-value is high (e.g., $p > 0.50$), the appropriate inference is that the null hypothesis is not rejected. As described by Badenes-Ribera et al. (2016), one of the most common misconceptions about *p*-values – even among trained researchers – is the "inverse probability fallacy," in which p-values are misinterpreted as the probability that the null hypothesis is true given the observed data (Carver, 1978). The

608 inverse probability fallacy can be formally expressed as p = p(H0|data). Consider the p-values
609 observed for the critical phase by ear interaction in the reaction time models for the full sample
610 in experiment 1 (p = 0.283) and experiment 2 (p = 0.634). Both p-values support the logical
611 inference that the null hypothesis is not rejected; however, neither of these p-values provide
612 direct support for the null hypothesis because, by definition, this is not the probability expressed
613 by the p-value.

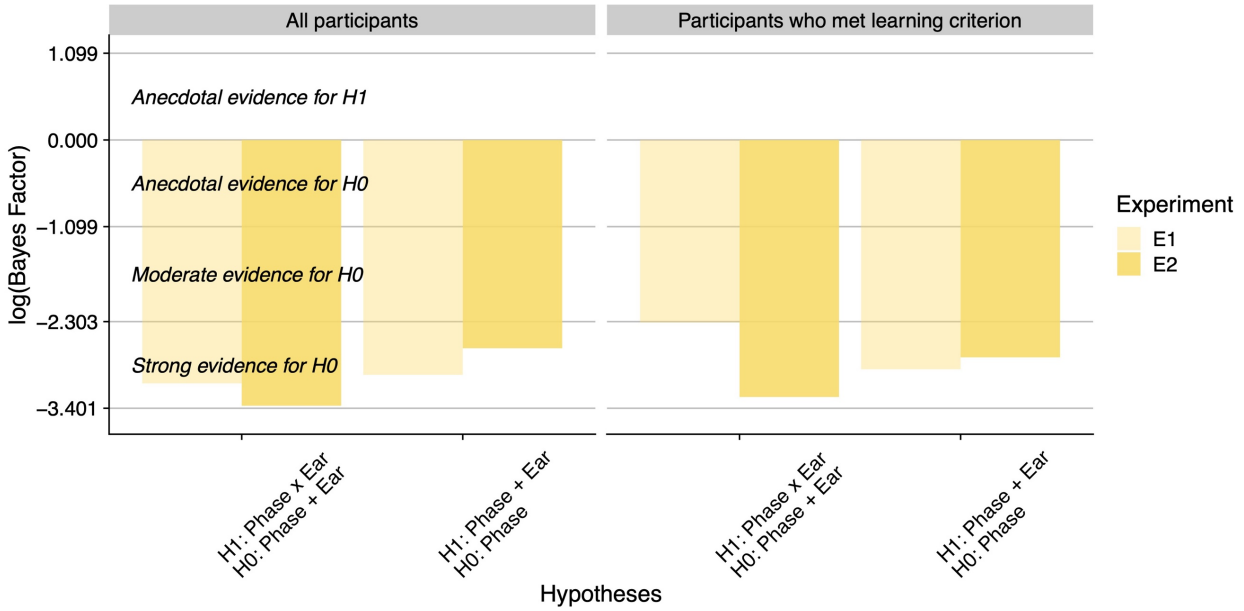
614 Bayes Factors analysis can help to resolve the inverse probability fallacy because the
615 Bayes Factor expresses the ratio between the likelihood of two hypotheses (e.g., Kass & Raftery,
616 1995; Lee & Wagenmakers, 2014; van Doorn et al., 2021). For example, a Bayes Factor can be
617 calculated for the likelihood of an alternative hypothesis (i.e., H1) relative to the likelihood of the
618 null hypothesis (i.e., H0), and thus can be interpreted as a measure of the strength of the evidence
619 in favor of one hypothesis over another. Unlike a p-value, the Bayes Factor can directly express
620 the strength of the evidence in support of the alternative *or* the null hypothesis. By convention, a
621 Bayes Factor of 1 is interpreted as no evidence for either hypothesis; that is, when the likelihood
622 of the H1 and H0 are equal, the Bayes Factor indicates no evidence for either hypothesis (e.g,
623 Lee & Wagenmakers, 2014). A Bayes Factor > 1 is interpreted as evidence in support of the H1
624 and a Bayes Factor < 1 is interpreted as evidence in support of the H0. Moreover, the magnitude
625 of the Bayes Factor can be interpreted as the degree of support. For example, Bayes Factors
626 between 3 and 10 are, by convention, interpreted as providing moderate evidence for the H1.
627 Likewise, Bayes Factors between 1/3 and 1/10 are interpreted as providing moderate evidence
628 for the H0. Accordingly, Bayes Factors analysis provides a tool for interpreting null effects
629 observed in frequentist analysis approaches because Bayes Factors support the interpretation of
630 null effects beyond the limited “failure to reject the null” inference that is licensed by p-values.

To this end, we calculated the Bayes Factor for the null effects that emerged in the reaction time models of experiments 1 and 2. All calculations were performed using the lmBF() function of the BayesFactor package (Morey & Rouder, 2018) in R, with multivariate Cauchy prior distributions set to scale = 0.5 and scale = 1.0 for fixed and random effects, respectively. Calculating a Bayes Factor requires specifying two hypotheses (i.e., models) for comparison, one to represent the H1 and one to represent the H0. For all calculations, we followed guidance to include random intercepts by subject and random slopes by subject for within-subjects variables when calculating Bayes Factors using trial-level data in mixed effects models; accordingly, null models reflected the balanced null (van Doorn et al., 2021).

Four Bayes Factors were calculated for each experiment, two for the model that included all participants and two for the model that included only the participants who met the learning criterion. First, we calculated the Bayes Factor when defining the H1 as a model that included fixed effects of phase, ear, and their interaction and the H0 as a model that only included fixed effects of phase and ear. The Bayes Factor here thus indicates the degree of support for the interaction versus the lack of interaction. The resulting Bayes Factors (on a natural log scale) are shown in Figure 5. In three of the four cases, the Bayes Factor indicated strong evidence in support of the H0 (i.e., no interaction between phase and ear); in the fourth case, the Bayes Factor was on the cusp between criteria used to mark moderate and strong support for the null hypothesis. Second, we calculated the Bayes Factor when defining the H1 as a model that included fixed effects of phase and ear and the H0 as a model that only included the fixed effect of phase. Accordingly, the Bayes Factor for these hypotheses indicates the degree of support for a model that includes ear as a predictor versus a model that does not. As shown in Figure 5, the resulting Bayes Factors indicated strong support for the null for both samples in each

experiment. Viewed in conjunction with the frequentist statistics reported for each experiment, the results of the Bayes Factors analysis suggest that each experiment provides strong support for the null hypothesis; that is, the current results support the hypothesis that ear of presentation does not influence reaction time in the current talker identification tasks.

FIG. 5. (Color online.) Bayes Factors analyses for the null effects observed in experiments 1 and 2. As described in the main text, Bayes Factors were calculated for two sets of hypotheses in each experiment, which were calculated separately for the full sample (i.e., all participants who met the a priori training criterion for inclusion in the study) and the subset of participants who met the learning criterion (defined as ≥ 5% improvement in talker identification accuracy from pre- to post-test). In the figure below, Bayes Factors are plotted on a natural log scale to facilitate visualization (i.e., a Bayes Factor of 1 = 0 on the natural log scale). Interpretation conventions are provided in italicized text, with gray lines indicating the bounds for each interpretation criterion (i.e., Bayes Factors between -2.303 and -3.401 on a natural log scale represent the range of values that can be interpreted as providing strong evidence in support of the null hypothesis).



V. DISCUSSION

Here we revisited the finding of Francis and Driscoll (2006), who showed that learning to use a phonetic property of speech as a cue to talker identity induced a left ear processing advantage for behavioral responses in a talker identification task. The left ear processing advantage was interpreted as evidence of hemispheric lateralization consistent with task

673 demands. As reviewed in the Introduction, this finding is broadly consistent with the
674 neuroimaging literature suggesting right hemisphere dominance for talker processing. However,
675 this finding is unexpected given the extant dichotic listening literature, which suggests that the
676 ability to measure hemispheric asymmetries through behavioral listening tasks requires
677 presenting competing stimuli across binaural channels. The current work aimed to (1) determine
678 whether a left ear advantage for phonetic cues to talker identification would generalize to a larger
679 sample and (2) identify factors that predict a listeners' ability to use phonetic cues for talker
680 identification. The results of the talker identification task converged across both experiments.
681 Specifically, listeners in the aggregate showed improved talker identification accuracy at post-
682 test compared to pre-test, indicative of learning to use VOT as a cue to talker identity, which did
683 support a behavioral processing advantage in terms of faster reaction times at post-test compared
684 to pre-test. However, we found no evidence to suggest a left ear advantage either in the full
685 sample (n = 97 in experiment 1, n = 79 in experiment 2) or in the subset of participants (n = 58 in
686 experiment 1, n = 55 in experiment 2) who met the Francis and Driscoll (2006) learning
687 criterion.

688 A failure to replicate could reflect any one of the methodological differences between the
689 original and current study. Some of these differences are more minor (e.g., different stimulus
690 sets, different number of training and test trials), whereas two differences are more substantial.
691 First, the current study used web-based measures for data collection and the original study tested
692 participants in a laboratory. It may be the case that the effect observed in Francis and Driscoll
693 (2006) may require a high level of control over the testing environment that is only possible in a
694 traditional laboratory setting. Though we cannot rule out this possibility, web-based and smart
695 phone-based methods have been shown to be sufficient for behavioral detection of cerebral

lateralization specifically (e.g., Bless et al., 2013; Parker et al., 2021) and for eliciting dichotic listening effects more generally (e.g., Milne et al., 2021; Woods et al., 2017). Gorilla Experiment Builder, the software used to deploy the current web-based study, provides excellent timing control for stimulus presentation and reaction time measurement (Anwyl-Irvine et al., 2021), and we followed best practice for web-based reaction times studies by implementing a fully within-subjects design so that differences in browser and hardware (which may influence experimental timing) were not confounded with experimental conditions. Moreover, all participants in the current study passed two dichotic listening tasks that were designed to screen for headphone compliance on web-based platforms (Milne et al., 2020; Woods et al., 2017) in addition to passing a custom channel detection screen.

Second, the talker identification task was completed in a single session in the current study, whereas the task was spread across three days in the original study. Accordingly, the left ear processing advantage in Francis and Driscoll (2006) may be linked to sleep-based consolidation (e.g., Earle et al., 2018). However, right hemisphere sensitivity to talker-specific VOT patterns for phonetic identification has been shown to emerge within an hour of exposure as measured using fMRI (Myers & Theodore, 2017), suggesting that neural sensitivity to a talker's phonetic signature is not contingent on sleep-based consolidation.

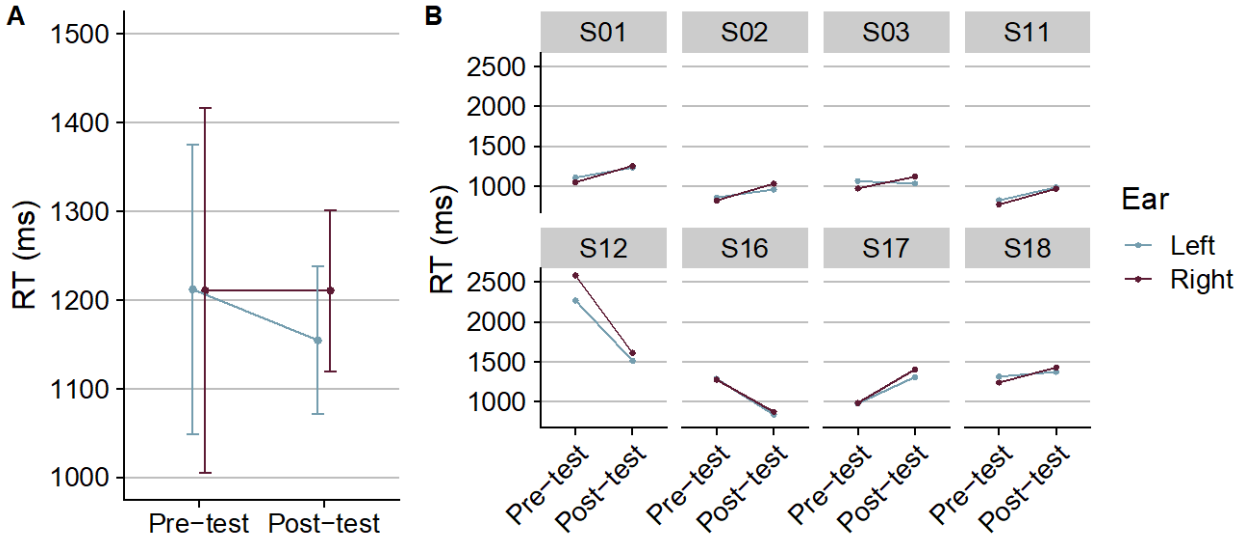
An additional explanation for the failure to replicate may be that the original effect reported in Francis and Driscoll (2006) was a false positive effect. This would not be unreasonable for three reasons. First, the paradigm was not optimized for measuring hemispheric laterality given the absence of a competing stimulus in the contralateral ear of interest (e.g., Kimura, 1967; Hugdahl, 2011). That is, although hemispheric laterality can be measured through behavioral dichotic listening tasks, Francis and Driscoll (2006) used a monaural listening task.

719 Monaural listening tasks show weakened ability to measure structural asymmetries of the
720 auditory pathway and its interaction with selective attention (Nicholls, 1998). In experiment 1,
721 we chose to replicate the original Francis and Driscoll task (i.e., present test stimuli monaurally)
722 to keep the replication closer to the original design. In experiment 2, pink noise was presented in
723 the contralateral ear to the target stimulus, which is a dichotic listening manipulation that has
724 successfully elicited a left ear advantage for talker identification in past research (González et al.,
725 2010). However, the left ear advantage failed to emerge in the current study even under these
726 more favorable conditions for behavioral observation of laterality effects. Second, the original
727 sample was limited to eight participants, and underpowered studies have been associated with an
728 increased rate of false positive effects (e.g., Button et al., 2013). Third, a reanalysis of the
729 Francis and Driscoll (2006) data suggests that the left ear effect was not stable at the level of
730 individual subjects, despite being significant in the aggregate. Figure 6 shows the aggregate
731 effect and by-subject patterns for each of the eight participants included in the aggregate
732 analysis. Five participants showed numerically faster mean reaction time for left compared to
733 right ear stimuli at post-test; however, none of the participants showed a pattern of responses
734 consistent with the group-level pattern.

735 Though we did not replicate a left ear advantage, the current results did show that
736 listeners could use VOT as a cue to talker identification, consistent with Francis and Driscoll
737 (2006) and numerous others studies pointing to tight links between the processing of phonetic
738 and indexical cues (e.g., Ganugapati & Theodore, 2019; Myers & Theodore, 2017; Theodore et
739 al., 2015; Theodore & Miller, 2010). Moreover, the current results shed light on individual
740 differences factors that predict listeners' use of VOT as a cue for talker identification.
741 Specifically, both measures of auditory acuity – pitch perception and within-category

discrimination – were positively associated with performance on the talker identification task; accuracy was higher at pre-test, training, and post-test for listeners with stronger vs. weaker auditory acuity. In contrast, inhibitory control and identification slope were not associated with any measure of talker identification performance. The only measure that predicted learning (i.e., the change in performance between pre- and post-test) was the location of participants' VOT voicing boundaries, with later boundaries associated with increased learning. Recall that Joanne and Sheila's characteristic VOTs were selected to reflect within-category variation for /k/. Participants with longer VOT voicing boundaries may have perceived the short VOT variants (i.e., Joanne) as members of the /g/ category, providing an additional cue for disassociating the talkers' voices.

FIG. 6. (Color online.) Re-analysis of Francis and Driscoll (2006). Panel A shows mean reaction time to correct responses at pre- and post-test by ear of presentation; error bars indicate standard error of the mean. Panel B shows performance for each of the eight participants who were included in the analysis presented in A; participant numbers (e.g., S01) reflect identifiers used in the original study.



On the one hand, null effects can present challenges for theory development (particularly when null effects emerge from frequentist analyses) because a failure to find evidence to reject the null hypothesis does not in turn provide evidence to support the null hypothesis. On the other

hand, observing cases where predictions of a given theory *do not hold* is a key component of the scientific method; these findings are needed to refine, revise, or perhaps even dismiss the theory. Moreover, any *single* test of a hypothesis cannot be considered definitive "truth;" this is only possible through repeated observations that together form a cumulative science. Though we imagine that these basic tenets of the scientific process are uncontroversial, these tenets are not reflected in the literature. For example, Scheel and colleagues (Scheel et al., 2021) examined the degree of hypothesis confirmation in the standard psychology literature compared to Registered Reports in this domain. A Registered Report is a relatively new research article type that is granted conditional acceptance prior to data collection; that is, the hypotheses and methods are evaluated independently from the results (e.g., Simons et al., 2014; Storkel & Gallun, 2022). They found that 96% of hypotheses were confirmed in the standard psychology literature compared to only 44% in Registered Reports. Concerns regarding the improbability of "successes" in the literature have been noted for decades (e.g., Fanelli, 2012; Scheel et al., 2021; Sterling, 1959; Sterling et al., 1995). These concerns have been linked to replication failures in numerous research domains (e.g., Begley & Ellis, 2012; Hubbard & Vetter, 1996; Ioannidis, 2005; Martin & Clarke, 2017), and some have argued the preponderance of positive effects in the literature is a direct consequence of a publication bias against null effects and replication studies (e.g., Neuliep, 1990; Neuliep & Crandall, 1993; Pashler & Wagenmakers, 2012). As argued by Haeffel (2022), in order to provide critical tests of our theories, we must stop the extraordinary "winning streak" that yields a scientific literature suggesting that positive support is obtained for every hypothesis that is tested.

If positive support for the hypothesis at hand is our metric of "winning," then we have definitely lost in the current study. However, given that the results of the Bayes Factors analyses

783 provided strong support for the null hypothesis, all is perhaps not lost for theory development.
784 The results of Francis and Driscoll (2006) led to the theory that hemispheric dominance reflects
785 functional use of acoustic-phonetic cues. That is, the theory posited that the *functional* use of a
786 given speech cue was the primary determinant of lateralization, and not the nature of the specific
787 speech cue. The current results are inconsistent with this theory. Combined with the extant
788 literature, they instead support a theory that places higher weight on the signal for guiding
789 hemispheric lateralization (Albouy et al., 2020; Perrachione et al., 2009; Poeppel, 2003; von
790 Kriegstein et al., 2003). For example, access to source characteristics may be necessary for
791 engaging right hemisphere dominance for voice processing. This is consistent with findings from
792 the two studies that observed behavioral evidence of right hemisphere lateralization for talker
793 identification. In González et al. (2010) and Perrachione et al. (2009), the stimuli consisted of
794 sentence-length items and talkers differed not only in their phonetic implementation of speech
795 sounds, presumably, but also in their indexical characteristics. That is, the talkers in these studies
796 differed on a host of naturally occurring dimensions, including fundamental frequency. As a
797 consequence, any talker-specific phonetic variability in the stimuli was conditioned on talker
798 differences in source characteristics. This was not the case in the current experiments because the
799 two talkers *only* differed in their phonetic instantiation of /k/.
800 Other behavioral research has shown that although talker-specific phonetic variation can
801 facilitate talker identification, listeners require additional time to learn the conditioning between
802 phonetic and indexical cues. For example, Ganugapati and Theodore (2019) trained listeners to
803 identify three female talkers from single-word utterances. For one group of listeners, phonetic
804 information was structured across talkers such that each talker had a characteristic VOT
805 production. This structure was absent for a different group of listeners who instead heard all

three talkers each produce all three characteristic VOTs. In contrast to the current experiments, the three talkers also differed in source characteristics; thus, sensitivity to talker-specific VOT was not required to perform the talker identification task. Indeed, given brief exposure to the talkers (72 trials), talker identification accuracy did not differ between the two groups. However, given longer exposure (216) trials, those who heard structured phonetic variation showed higher talker identification accuracy compared to those who did. Moreover, results from the neuroimaging literature suggest that right hemisphere regions associated with voice processing show sensitivity to talker-specific phonetic patterns when these patterns co-occur with talker differences in indexical cues (Myers and Theodore, 2017). Together with the current work, these findings are consistent with the theory that using talker-specific phonetic variation for voice processing will be heightened when phonetic variation can be conditioned on source characteristics. We note that though the current results are consistent with such a theory, future research is needed to confirm this hypothesis by direct examination of potential laterality effects following exposure to input that would allow phonetic variability to be conditioned on indexical variability.

In conclusion, the current study did not yield evidence to suggest a left ear processing advantage in behavior for the talker identification task used here. This does not imply that it may not emerge under different circumstances; indeed, the results contribute to a theory predicting that a right hemisphere (i.e., left ear) advantage may emerge when talker-specific phonetic variability can be conditioned on talker differences in source characteristics. Future research is needed to test this hypothesis directly. Given evidence to suggest right hemisphere sensitivity to talker-specific phonetic patterns (e.g., González et al., 2010; Myers & Theodore, 2017; Perrachione et al., 2009) and behavioral evidence indicating a tight coupling between phonetic

and indexical processing (e.g., Ganugapati & Theodore, 2019; Goggin et al., 1991; Nygaard & Pisoni, 1998; Orena et al., 2015), future research is warranted to determine the type and time course of behavioral advantages that may occur given perceptual learning of talker-specific phonetic detail.

ACKNOWLEDGEMENTS

This research was supported by National Science Foundation BCS grant 1827591 (RMT), National Institutes of Health grant NIH/NIDCD T32 DC017703, and by an Innovation Award from UConn's Neurobiology of Language program (LD). The views expressed here reflect those of the authors and not the National Science Foundation nor the National Institutes of Health. Portions of this project were completed as an Honors thesis by the second author under the direction of the fourth author.

REFERENCES

Albouy, P., Benjamin, L., Morillon, B., & Zatorre, R. J. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, *367*(6481), 1043–1047.

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *115*(6), 3171–3183.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544–552.

Anwyl-Irvine, A. L., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, *53*(4), 1407–1425.

851 Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our
852 midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388-
853 4071–20. https://doi.org/10.3758/s13428-019-01237-x
854 Badenes-Ribera, L., Frias-Navarro, D., Iotti, B., Bonilla-Campos, A., & Longobardi, C. (2016).
855 Misconceptions of the p-value among Chilean and Italian Academic Psychologists.
856 *Frontiers in Psychology*, *7*, 1247.
857 Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models
858 using lme4. *Journal of Statistical Software*, *67*(1), 1–48.
859 Bedford-Petersen, C. (2021). *Inauguration palette*. https://github.com/ciannabp/inauguration
860 Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*,
861 *483*(7391), 531–533.
862 Behne, N., Scheich, H., & Brechmann, A. (2005). Contralateral white noise selectively changes
863 right human auditory cortex activity caused by a FM-direction task. *Journal of*
864 *Neurophysiology*, *93*(1), 414–423.
865 Behne, N., Wendt, B., Scheich, H., & Brechmann, A. (2006). Contralateral white noise selectively
866 changes left human auditory cortex activity in a lexical decision task. *Journal of*
867 *Neurophysiology*, *95*(4), 2630–2637.
868 Bless, J. J., Westerhausen, R., Arciuli, J., Kompus, K., Gudmundsen, M., & Hugdahl, K. (2013).
869 “Right on all occasions?” On the feasibility of laterality research using a smartphone
870 dichotic listening application. *Frontiers in Psychology*, *4*, 42.
871 Bless, J. J., Westerhausen, R., Torkildsen, J. von K., Gudmundsen, M., Kompus, K., & Hugdahl, K.
872 (2015). Laterality across languages: Results from a global dichotic listening study using a

873 smartphone application. *Laterality: Asymmetries of Body, Brain and Cognition*, *20*(4),
874 434–452.
875 Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò,
876 M. R. (2013). Power failure: Why small sample size undermines the reliability of
877 neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.
878 https://doi.org/10.1038/nrn3475
879 Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*,
880 *48*(3), 378–399.
881 Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of
882 stop consonant VOT in American English. *Journal of Phonetics*, *61*, 30–47.
883 Earle, F. S., Landi, N., & Myers, E. B. (2018). Adults with Specific Language Impairment fail to
884 consolidate speech sounds during sleep. *Neuroscience Letters*, *666*, 58–63.
885 Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries.
886 *Scientometrics*, *90*(3), 891–904.
887 Francis, A. L., & Driscoll, C. (2006). Training to use voice onset time as a cue to talker
888 identification induces a left-ear/right-hemisphere processing advantage. *Brain and*
889 *Language*, *98*(3), 310–318.
890 Ganugapati, D., & Theodore, R. M. (2019). Structured phonetic variation facilitates talker
891 identification. *The Journal of the Acoustical Society of America*, *145*(6), EL469–EL475.
892 Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language
893 familiarity in voice identification. *Memory & Cognition*, *19*(5), 448–458.

González, J., Cervera-Crespo, T., & McLennan, C. T. (2010). Hemispheric differences in specificity effects in talker identification. *Attention, Perception, & Psychophysics*, *72*(8), 2265–2273.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498.

Haeffel, G. J. (2022). Psychology needs to get tired of winning. *Royal Society Open Science*, *9*(6), 220099.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111.

Hubbard, R., & Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, *18*(1), 69–88.

Hubbard, R., & Vetter, D. E. (1996). An empirical comparison of published replication research in accounting, economics, finance, management, and marketing. *Journal of Business Research*, *35*(2), 153–164.

Hugdahl, K., & Anderson, L. (1984). A dichotic listening study of differences in cerebral organization in dextral and sinistral subjects. *Cortex*, *20*(1), 135–141.

Ioannidis, J. P. (2005). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association*, *294*(2), 218–228.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

915 Kimura, D. (1967). Functional asymmetry of the brain in dichotic listening. *Cortex*, *3*(2), 163–
916 178.
917 Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear
918 mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.
919 https://doi.org/10.18637/jss.v082.i13
920 Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*.
921 Cambridge university press.
922 Long, J. A. (2020). *jtools: Analysis and presentation of social scientific data* (R package version
923 2.0.0). https://cran.r-project.org/package=jtools
924 Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of
925 editorial practices. *Frontiers in Psychology*, *8*, 523.
926 Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An
927 online headphone screening test based on dichotic pitch. *Behavior Research Methods*,
928 *53*(4), 1551–1562.
929 Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for common*
930 *designs* (0.9.12-4.3).
931 Myers, E. B., & Theodore, R. M. (2017). Voice-sensitive brain networks encode talker-specific
932 phonetic detail. *Brain and Language*, *165*, 33–44.
933 https://doi.org/10.1016/j.bandl.2016.11.001
934 Neuliep, J. W. (1990). Editorial bias against replication research. *Journal of Social Behavior and*
935 *Personality*, *5*(4), 85.

936 Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of*
937 *Social Behavior and Personality*, *8*(6), 21.
938 Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Attention,*
939 *Perception, & Psychophysics*, *60*(3), 355–376.
940 Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent
941 process. *Psychological Science*, *5*(1), 42–46.
942 Orena, A. J., Theodore, R. M., & Polka, L. (2015). Language exposure facilitates talker learning
943 prior to language comprehension, even in adults. *Cognition*, *143*, 36–40.
944 Palan, S., & Schitter, C. (2018). Prolific. Ac—A subject pool for online experiments. *Journal of*
945 *Behavioral and Experimental Finance*, *17*, 22–27.
946 Parker, A. J., Woodhead, Z. V., Thompson, P. A., & Bishop, D. V. (2021). Assessing the reliability
947 of an online behavioural laterality battery: A pre-registered study. *Laterality:*
948 *Asymmetries of Body, Brain and Cognition*, *26*(4), 359–397.
949 Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on
950 replicability in psychological science: A crisis of confidence? *Perspectives on*
951 *Psychological Science*, *7*(6), 528–530.
952 Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. (2011). Human voice recognition depends on
953 language ability. *Science*, *333*(6042), 595–595.
954 Perrachione, T. K., Pierrehumbert, J. B., & Wong, P. (2009). Differential neural contributions to
955 native-and foreign-language talker identification. *Journal of Experimental Psychology:*
956 *Human Perception and Performance*, *35*(6), 1950.

957 Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral
958 lateralization as 'asymmetric sampling in time.' *Speech Communication*, *41*(1), 245–255.
959 Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the
960 standard Psychology literature with Registered Reports. *Advances in Methods and*
961 *Practices in Psychological Science*, *4*(2), 1–12.
962 Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered
963 replication reports at Perspectives on Psychological Science. *Perspectives on*
964 *Psychological Science*, *9*(5), 552–555.
965 Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design.
966 *Psychonomic Bulletin & Review*, *25*(6), 2083–2101.
967 Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from
968 tests of significance—Or vice versa. *Journal of the American Statistical Association*,
969 *54*(285), 30–34.
970 Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The
971 effect of the outcome of statistical tests on the decision to publish and vice versa. *The*
972 *American Statistician*, *49*(1), 108–112.
973 Storkel, H. L., & Gallun, F. J. (2022). Announcing a new registered report article type at the
974 Journal of Speech, Language, and Hearing Research. *Journal of Speech, Language, and*
975 *Hearing Research*, *65*(1), 1–4.
976 Studdert-Kennedy, M., & Shankweiler, D. (1970). Hemispheric specialization for speech
977 perception. *The Journal of the Acoustical Society of America*, *48*(2B), 579–594.

978 Theodore, R. M., & Flanagan, E. G. (2020). Determinants of voice recognition in monolingual
979 and bilingual listeners. *Bilingualism: Language and Cognition*, *23*(1), 158–170.
980 Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific
981 phonetic detail. *The Journal of the Acoustical Society of America*, *128*(4), 2090–2099.
982 Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-
983 time: Contextual influences. *The Journal of the Acoustical Society of America*, *125*(6),
984 3974–3982. https://doi.org/10.1121/1.3106131
985 Theodore, R. M., Myers, E. B., & Lomibao, J. A. (2015). Talker-specific influences on phonetic
986 category structure. *The Journal of the Acoustical Society of America*, *138*(2), 1068–1078.
987 van Doorn, J., Aust, F., Haaf, J. M., Stefan, A. M., & Wagenmakers, E.-J. (2021). Bayes factors for
988 mixed models. *Computational Brain & Behavior*, 1–13. https://doi.org/10.1007/s42113-
989 021-00113-2
990 von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural
991 responses to speech by directing attention to voices or verbal content. *Cognitive Brain*
992 *Research*, *17*(1), 48–55.
993 Westerhausen, R. (2019). A primer on dichotic listening as a paradigm for the assessment of
994 hemispheric asymmetry. *Laterality: Asymmetries of Body, Brain and Cognition*, *24*(6),
995 740–771.
996 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G.,
997 Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the tidyverse. *Journal of Open*
998 *Source Software*, *4*(43), 1686.

Wilke, C. O. (2019). *cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2"* (R package version 0.9.4). https://CRAN.R-project.org/package=cowplot

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072.

Xie, X., & Myers, E. (2015). The impact of musical training and tone language experience on talker identification. *The Journal of the Acoustical Society of America*, *137*(1), 419–432. https://doi.org/10.1121/1.4904699