Relationships between native and non-native speech perception

Pamela Fuhrmeister[1], Matthew C. Phillips[1], D. Betsy McCoach[2], & Emily B. Myers[1]

[1] University of Connecticut, Department of Speech, Language, and Hearing Sciences

[2] University of Connecticut, Neag School of Education

Author Note

This material is based upon work supported in part by the National Science Foundation under grant DGE-1747486 to the University of Connecticut and NSF BCS 1554510 to EBM. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors are grateful to Hannah Mechtenberg for help with data collection. The data and analysis code for the experiments in this paper can be found at https://osf.io/2zg6c/.

Correspondence concerning this article should be addressed to Pamela Fuhrmeister. E-mail: pamela.fuhrmeister@uconn.edu

Abstract

Individuals differ in their ability to perceive and learn unfamiliar speech sounds, but we lack a comprehensive theoretical account that predicts individual differences in this skill. Predominant theories largely attribute difficulties of non-native speech perception to the relationships between non-native speech sounds/contrasts and native-language categories. The goal of the current study was to test whether the predictions made by these theories can be extended to predict individual differences in naive perception of non-native speech sounds or learning of these sounds. Specifically, we hypothesized that the internal structure of native-language speech categories is the cause of difficulty in perception of unfamiliar sounds such that learners who show more graded (i.e., less categorical) perception of sounds in their native language would have an advantage for perceiving non-native speech sounds because they would be less likely to assimilate unfamiliar speech tokens to their native-language categories. We tested this prediction in two experiments in which listeners categorized speech continua in their native language and performed tasks of discrimination or identification of difficult non-native speech sound contrasts. Overall, results did not support the hypothesis that individual differences in categorical perception of *native-language* speech sounds is responsible for variability in sensitivity to *non-native* speech sounds. However, participants who responded more consistently on a speech categorization task showed more accurate perception of non-native speech sounds. This suggests that individual differences in non-native speech perception are more related to the stability of phonetic processing abilities than to individual differences in phonetic category structure.

Keywords: speech perception, categorical perception, individual differences, non-native speech sound learning, perceptual assimilation

Relationships between native and non-native speech perception

Many adults struggle to perceive and produce speech sounds in a second language (e.g., Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999). Some obvious factors contribute to more successful learning of speech sounds, including the age at which someone starts learning the language or the amount of time they spend using their second language (e.g., Flege, Yeni-Komshian, & Liu, 1999; Piske, MacKay, & Flege, 2001). However, even among learners with similar experience with a second language, we still see a large range of individual variability in perception and production of non-native speech sounds: Some learners meet with excellent success in distinguishing difficult non-native contrasts, where others never manage to grasp the distinctions (e.g., Bradlow et al., 1999; Golestani & Zatorre, 2009; Lim & Holt, 2011; Luthra et al., 2019). What, then, accounts for this variability? A collection of findings suggests that perception or production of non-native speech sounds may be related to an individual's cue-weighting strategies or abilities (Schertz, Cho, Lotto, & Warner, 2015), auditory processing skills (Kachlicka, Saito, & Tierney, 2019), musical training/pitch perception (especially for learning of tonal contrasts, Bowles, Chang, & Karuzis, 2016; Kempe, Bublitz, & Brooks, 2015; Perrachione, Lee, Ha, & Wong, 2011; Slevc & Miyake, 2006), phonological skills (Earle & Arthur, 2017; Fuhrmeister, Schlemmer, & Myers, 2020; Perrachione et al., 2011), or brain structure (Golestani, Molko, Dehaene, LeBihan, & Pallier, 2007; Golestani, Paus, & Zatorre, 2002).

In the current study, we test a prediction that falls out of dominant theories of non-native speech perception: namely that the challenge of non-native speech perception is the result of the difficulties that adults have in detecting differences between items within native language speech categories, a phenomenon known as "categorical perception." As will be described in detail in the following paragraphs, popular theories of non-native speech perception predict that the relationship between native and non-native speech sounds explains the difficulty in perceiving non-native sounds (e.g., Best, McRoberts, &

Goodell, 2001; Best & Tyler, 2007; Flege, 1995; Iverson et al., 2003; Kuhl, 1994; Kuhl et al., 2008). These theories reliably predict *which* speech sounds or speech sound contrasts will be relatively easier or relatively difficult to perceive for a listener of a given native-language based on the relationship between native and non-native speech sounds (e.g., Best et al., 2001; Best, McRoberts, & Sithole, 1988). However, it is unknown whether these theories can be extended to predict which *individuals* will be more successful non-native speech sound perceivers or learners. Silbert et al. (2015) tested whether individuals have a general aptitude for perceiving non-native speech sound contrasts, and instead of a general aptitude, their results indicate that listeners tended to show different patterns of perception for segmental vs. tonal contrasts. This suggests that we may need to consider relationships between various skills and perception of segmental and suprasegmental contrasts separately, and hints that difficulty perceiving non-native contrasts might be tied to more specific relationships between an individual non-native contrast and corresponding native category representations. The goal of the current study is to test the hypothesis that individual differences in native-language speech category representations (specifically how categorically individuals perceive sounds) can predict success in perceiving or learning new speech categories. In the following paragraphs, we discuss categorical perception of speech, predictions from theories of non-native speech perception, and our reasoning for bringing these literatures together to generate new predictions for individual differences in non-native speech perception.

Individual differences in categorical perception of native-language speech sounds

For some time, the dominant view in the field has been that listeners perceive speech sounds categorically, i.e., acoustic differences between speech categories are highly detectable, while differences within speech categories are often poorly detected (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; e.g., Liberman, Harris, Hoffman, &

Griffith, 1957). Since this seminal finding, much research in the field has expanded on this phenomenon but has only recently begun to examine whether individuals differ systematically in whether their perception of phonetic categories is more categorical or more graded (e.g., Kapnoula, Winn, Kong, Edwards, & McMurray, 2017; Kong & Edwards, 2016).

One potential limitation of previous studies is that two-alternative forced choice identification tasks that are often used to measure categorical perception do not allow listeners to sufficiently demonstrate gradedness in their perception of speech sounds. For instance, when only given the choice between /d/ and /t/, many listeners who could otherwise detect subtle acoustic differences between two exemplars in the /d/ category would still identify all or most /d/ tokens as belonging to the /d/ category because those tokens would still be better exemplars of /d/ than /t/. Other tasks for measuring categorical perception offer the listener more opportunities to demonstrate graded perception of speech sounds. Some of these methods include eye-tracking (e.g., Clayards, Tanenhaus, Aslin, & Jacobs, 2008; McMurray, Danelz, Rigler, & Seedorff, 2018; McMurray, Tanenhaus, & Aslin, 2002), goodness judgments (Drouin, Theodore, & Myers, 2016; Miller, 1997), or visual analogue scales, in which a listener moves a visual slider on a screen between two alternatives (Kapnoula, Edwards, & McMurray, 2021; Kapnoula et al., 2017; Kong & Edwards, 2016). These studies suggest that when the task affords an opportunity to demonstrate sensitivity to variation within the phonetic category, listeners show sensitivity to the graded internal structure of speech categories (Clayards et al., 2008; Drouin et al., 2016; McMurray et al., 2018, 2002; Miller, 1997). Furthermore, considerable variability exists even among typically developing individuals in how graded or categorically speech categories are represented when measured in this way (Kapnoula et al., 2021, 2017; Kong & Edwards, 2016). Taken together, these findings suggest that while most listeners do continue to show a classic "categorical perception" pattern, with lessened sensitivity to sound contrasts falling within a category, these listeners differ systematically

in the degree to which they show a more categorical vs. graded perceptual pattern. Far from having uniform perceptual abilities, adults differ in the degree to which they can detect fine-grained phonetic differences within a native-language speech category. Given that theoretical accounts of non-native speech perception (discussed below) hinge on assimilation within the native category, this makes *native-language* categoricity/gradiency an attractive explanation for individual differences in the difficulty of *non-native* speech perception.

Theoretical predictions from non-native speech sound learning

Several theories attribute difficulties in non-native speech sound learning to perceptual similarity of non-native contrasts to native language speech categories (e.g., Best & Tyler, 2007; Flege, 1995; Iverson et al., 2003; Kuhl et al., 2008). For example, the native-language magnet model discusses this in terms of category prototypes: native-language categories will act as a "magnet" so that unfamiliar speech sounds that are acoustically or perceptually similar to native-language sounds will often be perceived as exemplars of the native-language category that is more familiar to the listener (Kuhl, 1994; Kuhl et al., 2008). Similarly, Iverson et al. (2003) provide a perceptual interference account of non-native speech sound learning. They suggest that experience with a (native) language changes perception and perhaps also lower-level auditory processing for certain acoustic dimensions, which in turn interferes with learning new sounds that are contrasted by that dimension.

The perceptual assimilation model predicts that non-native speech contrasts will be perceptually assimilated to native-language sounds, and these assimilation patterns determine how difficult it will be for a listener to differentiate these non-native speech sounds (Best et al., 2001; Best & Tyler, 2007). Specifically, the distinctions between two non-native speech sounds that assimilate to one native-language speech category will be particularly difficult to perceive, whereas contrasts that assimilate to two different native

speech categories will be easier (Best et al., 2001). For example, the voiced dental and retroflex stop consonants in Hindi are often difficult for native speakers of many varieties of English to differentiate because they are allophones of the alveolar /d/ category (e.g., in "width" or "address," Polka, 1991). These sounds then get assimilated to the alveolar category, making it difficult to perceive them as separate sounds. Mayr and Escudero (2010) found tentative support for the idea that individual differences in assimilation patterns between native and non-native speech sounds predicted more accurate perception of non-native contrasts; however, Hattori and Iverson (2009) tested a similar question but did not find much support for the notion that the degree of assimilation between a difficult non-native contrast and the perceptually similar native-language category was related to perception or production accuracy of the non-native contrast.

The speech learning model also assumes that the (dis)similarity between phonemes in the first and second language will influence the likelihood of establishing new phonological categories in a second language (Flege, 1995). This model focuses on experienced second-language learners rather than naive listeners and therefore a longer time span of learning; however, many of the core ideas and predictions are similar to the other models discussed here.

Although these theories are highly accurate in predicting *which* non-native speech sounds will be most difficult to acquire for learners of a given language background, they do not explain the variability commonly observed among learners with the *same* native language. Notably, these models do not make explicit predictions about individual differences in non-native speech perception, but implicit in the models' assumptions is that variation in the perception of native-language speech category structure should predict how easy or difficult acquiring perceptual sensitivity to a new, non-native speech sound contrast will be for an individual. Therefore, these theories may be able to generate predictions on an individual level: All of them would likely predict a relationship between subtle variation

in how native-language categories are perceived and how well someone can perceive or learn non-native speech sounds. For instance, if an individual shows perception of native-language categories that is more graded and less categorical (i.e., that person can distinguish subtle within-category differences), they may be less likely to assimilate similar-sounding non-native speech sounds to existing native categories. In other words, those with more graded native-language categories may be better equipped to detect acoustic variability that defines non-native contrasts.

Current study

The primary aim of the current study is to test whether individual differences in categorical perception of native-language speech sounds predict success on non-native speech sound learning tasks. We used a visual analogue scale task similar to those used in Kapnoula et al. (2017) and Kong and Edwards (2016) to derive a continuous measure of how categorically or graded participants perceive native-language speech sounds, which we refer to as *categoricity*. A finding in which categoricity negatively predicts non-native speech sound learning outcomes (i.e., participants who are less categorical are more successful on non-native tasks) would lend support to the idea that theories of non-native speech sound learning, such as the native-language magnet model (Kuhl, 1994; Kuhl et al., 2008), perceptual interference model (Iverson et al., 2003), perceptual assimilation model (Best et al., 2001; Best & Tyler, 2007), or speech learning model (Flege, 1995) can be extended to account for individual differences in non-native speech sound learning. In the current study we used a discrete visual analogue scale that had seven discrete points on the line (one for each continuum point). This allowed us to measure not only categoricity, but also how consistently participants rated the stimuli each time they were presented (i.e., how often they rated the first step on the continuum as "1" when it was presented). Some recent evidence suggests that perception of native-language speech sounds becomes more graded throughout adolescence (McMurray et al., 2018) and that adult speech category

representations are more graded than might be reflected by a two-alternative forced choice task McMurray et al. (2002). McMurray et al. (2018) argue that shallower categorization slopes in two-alternative forced choice phonetic identification tasks are a result of noisy representations, rather than the older view that they are due to more graded representations (e.g., Burnham, Earnshaw, & Clark, 1991; Hazan & Barrett, 2000). Therefore, we also collected a measure of how precise (or noisy) a participant's responses on the task were (*response consistency*). If more precise representations of speech sounds allow for the integration and formation of new speech categories, we would expect to see that more consistent responses on this task predict non-native speech sound learning.

Experiment 1

In Experiment 1, we tested whether categoricity and response consistency of two different native-language speech contrasts (a stop and a fricative contrast) predicted naive discrimination of a difficult non-native speech contrast or learning/retention of this contrast after a brief training period.

Method

Participants

Fifty-eight participants were recruited from the University of Connecticut (Connecticut, United States) community. All participants were native speakers of North American English and had typical hearing and no reading or language disorders. One participant could not complete the experiment due to an equipment error; data from the remaining 57 participants are included in the analyses. All participants gave informed consent according to the University of Connecticut Institutional Review Board requirements and were paid $10 per hour for participation.

Stimuli

We tested participants' discrimination and identification of a difficult non-native speech sound contrast for our participants (native speakers of North American English), the voiced dental and retroflex stop consonants found in Hindi. We recorded five exemplars each of the minimal pair non-words /d̪ug/ and /d̪ug/. Stimuli were recorded by a female native speaker of Hindi; stimuli were recorded in a soundproof booth and were scaled to a mean amplitude of 65 dB sound pressure level in Praat (Boersma & Weenink, 2013). Fribbles served as novel objects to pair new words with (stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, http://www.tarrlab.org/). Non-native tasks were presented on a desktop computer with the experiment presentation software OpenSesame (Mathôt, Schreij, & Theeuwes, 2012).

To measure perception of native-language speech categories, participants rated tokens from two seven-step continua on the visual analogue scale. One continuum consisted of a fricative contrast embedded in real words (sign-shine) and one consisted of a synthetic stop contrast of consonant-vowel syllables (/bɑ/-/dɑ/)[1]. The /bɑ/-/dɑ/ continuum was created using a Klatt synthesizer at Haskins Laboratories. Stimuli for the sign-shine continuum were recorded by a female, native speaker of English, and waveform averaging in Praat (Boersma & Weenink, 2013) was used to create blends from 20% /s/ to 80% /ʃ/ in 10% steps. Native-language visual analogue scale tasks were presented using E-Prime 3.0 (Psychology Software Tools, Pittsburgh, PA). We chose these two contrasts because we did not want the acoustic properties of the stimuli to be too different between the native and non-native speech tasks. For instance, dental and retroflex stop consonants are

[1] We did not have strong theoretical reasons for including contrasts embedded in both real words and syllables other than that we know that these are well validated continua that elicit typical categorical patterns of perception.

differentiated by the frequency of the burst and the third and fourth formant trajectories, i.e., primarily spectral cues (Stevens & Blumstein, 1975). We therefore chose two native-language contrasts that are also mainly differentiated by spectral cues.

Procedure

Participants came to the lab twice. The first session took place in the evening hours (between 5 and 9 PM) and included consent and non-native phonetic training and assessments in this order: discrimination pretest[2], identification training, identification posttest, and discrimination posttest. The second session took place the following morning between 8 and 10 AM and included (in this order) a test of retention of non-native speech sounds (identification posttest followed by the discrimination posttest) and after that, the visual analogue scale tasks to measure categorical perception of native-language speech sounds (the order of the continua was counterbalanced)[3]. The training and testing schedule was motivated by previous work showing that relationships between phonological skills in the native-language and a non-native speech sound learning task emerged after a period of sleep (Earle & Arthur, 2017). We therefore wanted to test the possibility that relationships between categoricity or consistency and non-native speech perception tasks would emerge only after a period of offline consolidation.

Non-native speech sound learning tasks.

Discrimination. The discrimination task was an AX task (i.e., same/different judgment). On each trial (64 trials total), participants heard two of the minimal pair non-words with an inter-stimulus interval of 1 s and were asked to indicate whether the non-words started with the same or different sounds. Four practice trials with feedback

[2] We chose to only administer a discrimination pretest and not an identification pretest. In similar study designs, we have never done a pretest for an identification task because that requires at least some amount of familiarization with the stimuli. A discrimination task, on the other hand, can be done naively, as participants simply indicate whether they think the sounds they heard were the same or different.

[3] We additionally collected a battery of tests of cognitive skills; these will be reported elsewhere.

were administered, and these are not included in the analyses. No feedback was given on any subsequent trials.

Identification. First, participants were familiarized with the non-word-object pairings by seeing one visual stimulus (Fribble) on the screen at a time while hearing all five auditory exemplars of that non-word. Next, participants learned the sounds via a two-alternative forced choice identification task. For this task, they saw both objects on the screen and heard one auditory token and then were asked to choose the corresponding picture. Visual feedback ("Correct"/"Incorrect") was given on each trial. Participants completed a total of 400 trials with a two-minute break halfway through. After training, participants completed an identification assessment, which included 50 trials just like in training but with no feedback.

Native-language speech perception measures. Participants completed two tasks measuring categoricity of native-language speech sounds. In each task, auditory tokens taken from one of two continua were presented (sign-shine or ba-da), and the order of the two continua was counterbalanced. Participants indicated their response on a visual analogue scale. In this task, participants moved a slider to discrete (numbered) points on the line between two stimuli (in this case, words/syllables) to indicate how, for example, sign-like or shine-like each stimulus token sounded (see Figure 1 for an illustration of the task and Figure 2 for examples of a graded and categorical participant).

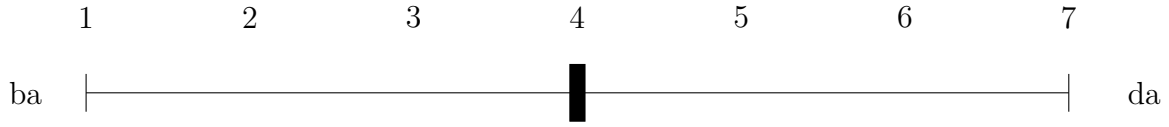

Figure 1. Illustration of the discrete visual analogue scale task.

Analysis approach

All analyses were performed in R (R Core Team, 2021). Categoricity and non-native discrimination data were analyzed with generalized linear mixed effects models with the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). Unless otherwise noted, the full

random effects structure was used unless the model did not converge, in which case, random effects were removed one by one until the model converged without warnings (Barr, Levy, Scheepers, & Tily, 2013). All data and code to reproduce the analyses reported in this paper can be found at our OSF repository: https://osf.io/2zg6c/.

Non-native discrimination. A common way to analyze data from a discrimination task such as the one used here is to compute a d prime (d') score to account for response bias (Macmillan & Creelman, 2004). For instance, when indicating whether stimuli sound the same or different, a participant who responds "different" on every trial in the experiment would score 100% for all the different trials. However, this does not mean that the participant can actually distinguish the sounds because they also answered incorrectly for all the same trials. Using d' scores can overcome this bias by subtracting the standardized "false alarm" score ("same" trials on which the participant answered "different") from the standardized "hit" score ("different" trials on which the participant correctly answered "different"), i.e., z(hits)-z(false alarms). In short, this score represents the distance between the standardized distributions of responses when the stimuli are different and when they are the same. However, this approach requires averaging over trials, and we would rather incorporate this approach into a hierarchical model to take trial and item variability into account while still accounting for response bias. This can be done with a generalized linear mixed effects model with the probit link because the probit link transformation is the inverse of the cumulative distribution function of the standard normal distribution (Razzaghi, 2013; Wright, Horry, & Skagerberg, 2009).

To test whether categoricity or response consistency predicted accuracy on the non-native discrimination task, we fit a generalized linear mixed effects model with the probit link. The model predicted whether the response on each trial was different (yes = 1, no = 0). Fixed effects included condition (deviation coded: different = .5, same = -.5), the interaction of condition and time (backwards difference coded to test contrasts for learning: immediate posttest-pretest, and retention: next-day posttest-immediate posttest), and the

interaction of condition and each predictor variable (ba-da categoricity, ba-da consistency, s-sh categoricity, and s-sh consistency, all mean centered) was nested within time. Nesting the predictors within the factor of time tests for simple effects of the predictors at each time point separately (Schad, Vasishth, Hohenstein, & Kliegl, 2020, e.g., whether categoricity predicted discrimination accuracy at the pretest, immediate posttest, or next-day posttest separately). In this model, the intercept represents overall bias, or whether there was more of a tendency to answer "same" or "different" regardless of condition. The coefficient for condition represents d', or sensitivity considering the actual condition (same/different). The interactions are what we are most interested in for the present question, which can be interpreted as the change in sensitivity (d') for each unit increase in categoricity or response consistency. We did not estimate main effects of categoricity or response consistency because they would not be informative for our question without interpreting them in the context of the interaction with condition.

Non-native identification. With a two-alternative forced choice identification task without feedback, we sometimes see that participants score well below chance performance (for example scoring 2% accuracy on the task). This makes sense if participants can indeed distinguish the sounds but they simply switched the labels. To address this, we recoded correct and incorrect responses for identification data (i.e., 0 was recoded as 1 and 1 was recoded as 0) if participants' mean accuracy was significantly below chance as measured by a binomial test (below 38% accuracy for 50 trials). No participants scored below chance, so no data was recoded in this experiment.

We analyzed identification performance using a generalized linear mixed effects model with a logit link that predicted accuracy (0 or 1). Fixed effects included ba-da categoricity, ba-da consistency, s-sh categoricity, and s-sh consistency (all mean centered), which were nested within time (deviation coded: immediate posttest = -.5, next-day = .5).

Native-language speech perception measures. To measure categoricity of native-language speech sounds, we fit responses from the visual analogue scale (for each

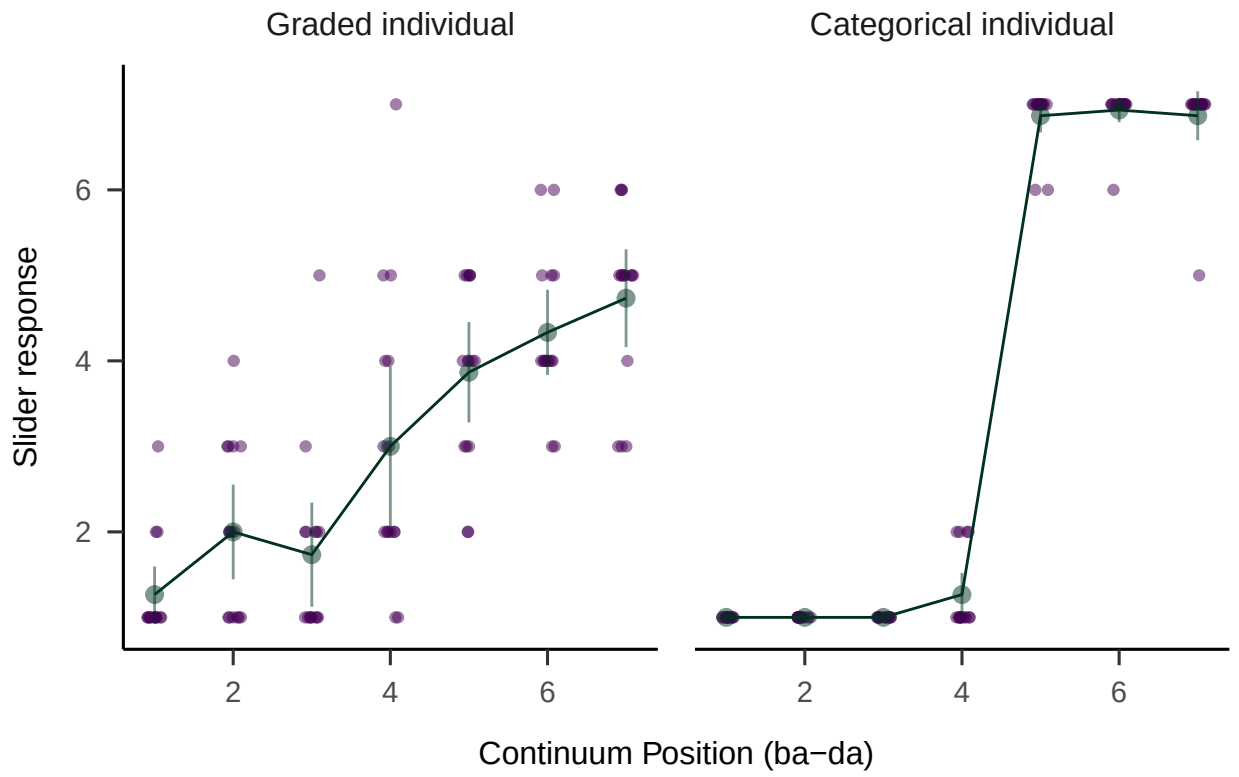

Figure 2. Examples of participants who showed more graded (left) and categorical (right) patterns of perception on the task with the visual analogue scale.

continuum separately) to a three-parameter logistic function with a non-linear mixed effects regression model. The three-parameter model estimates the maximum asymptote, the inflection point (i.e., category boundary), and the slope of the function (more categorical response patterns have higher slope values). Non-linear regression models were fit with the nlme package (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2021). Each participant's categoricity score (the population mean plus the participant's individual random effect adjustment) was extracted from the model. An individual measure of response consistency on each phonetic continuum was obtained by calculating the mean of the squared residuals from the model for each participant separately (residuals were squared to avoid negative values). Because larger values represented less consistent responses, we changed the sign of the response consistency measure for ease of interpretation, such that *larger* values represent *more* consistent response patterns. The measures of slope and consistency were

entered into further analyses described below. We additionally computed all the possible correlations between these measures. The correlation between the categoricity measures from the two continua were moderately correlated, $r = 0.31$, $p = 0.02$; all other correlations were non-significant: consistency for both continua, $r = 0.16$, $p = 0.22$, ba-da categoricity and consistency: $r = -0.07$, $p = 0.60$, s-sh categoricity and consistency: $r = -0.05$, $p = 0.74$.

Table 1

Results of the discrimination task in Experiment 1. Time1 = pretest, Time2 = immediate posttest, and Time3 = next-day posttest.

Predictors	*Estimate*	*SE*	*z*	*p*
Intercept	-0.07	0.05	-1.52	0.13
Condition (d')	1.26	0.15	8.39	<.001
Time2-1	0.02	0.03	0.65	0.51
Time3-2	0.00	0.03	0.01	0.99
Time2-1 x Condition	0.64	0.07	9.84	<.001
Time3-2 x Condition	0.19	0.07	2.85	<.001
Time1:Condition x Categoricity ba-da	-0.17	0.43	-0.39	0.7
Time2:Condition x Categoricity ba-da	0.41	0.43	0.95	0.34
Time3:Condition x Categoricity ba-da	0.15	0.43	0.34	0.73
Time1:Condition x Consistency ba-da	0.35	0.35	0.99	0.32
Time2:Condition x Consistency ba-da	0.00	0.36	-0.01	1
Time3:Condition x Consistency ba-da	-0.11	0.36	-0.30	0.77
Time1:Condition x Categoricity s-sh	0.11	1.04	0.10	0.92
Time2:Condition x Categoricity s-sh	-1.28	1.05	-1.22	0.22
Time3:Condition x Categoricity s-sh	-1.23	1.05	-1.17	0.24
Time1:Condition x Consistency s-sh	0.58	0.33	1.73	0.08
Time2:Condition x Consistency s-sh	0.42	0.34	1.25	0.21

Time3:Condition x Consistency s-sh	0.31	0.34	0.92	0.36

Table 2

Results of the identification task in Experiment 1. Time1 = immediate posttest, and Time2 = next-day posttest.

Predictors	*Estimate*	*SE*	*z*	*p*
Intercept	2.30	0.26	8.80	<.001
Time	0.13	0.12	1.10	0.27
Time1:Categoricity ba-da	1.07	0.62	1.74	0.08
Time2:Categoricity ba-da	0.17	0.67	0.25	0.81
Time1:Categoricity s-sh	-2.76	1.47	-1.88	0.06
Time2:Categoricity s-sh	-2.31	1.64	-1.41	0.16
Time1:Consistency ba-da	-0.36	0.50	-0.72	0.47
Time2:Consistency ba-da	-0.33	0.56	-0.60	0.55
Time1:Consistency s-sh	0.92	0.47	1.97	0.05
Time2:Consistency s-sh	0.73	0.52	1.39	0.16

Results

Discrimination

Results of the discrimination analysis are summarized in Table 1. We found a significant main effect of condition, meaning participants answered "different" more often when the stimulus pair was different. There were significant interactions of condition with each of the time contrasts, which indicates that participants improved in their discrimination accuracy of the Hindi sounds at each time point. None of the measures of categoricity or consistency significantly predicted discrimination accuracy.

366 ### Identification

367 Results of the identification analysis are summarized in Table 2. The intercept was
368 significantly different from zero, indicating participants identified the sounds above chance,
369 but there were no other significant effects[4]. The relationship between consistency on the
370 fricative continuum and identification accuracy approached significance at the immediate
371 posttest. This is also in the expected direction, in that more consistent responses predicted
372 higher accuracy; however, we did not see the same pattern with the stop continuum, so we
373 will err on the side of caution and not interpret this further.

374 ## Discussion

375 Neither categoricity nor consistency significantly predicted discrimination or
376 identification accuracy at any of the time points we tested. One obvious explanation is that
377 we lacked statistical power in this design. We address this issue in Experiment 2.
378 Additionally, if independent variables in a regression model are highly correlated, we may
379 not see any significant effects because no one variable explains unique variance. We only
380 saw weak correlations (if any) among predictor variables of categoricity and consistency;
381 however, to ensure that our null results were not due to issues of multicollinearity, we
382 calculated variance inflation factors (a diagnostic tool to detect multicollinearity in
383 multiple regression models) for all predictors in the model using the car package in R (Fox
384 & Weisberg, 2019). Variance inflation factors of over 5 (or over 10 according to some
385 authors) are considered problematic (e.g., Thompson, Kim, Aloe, & Becker, 2017). Most
386 variance inflation factors were around 1 with only a couple closer to 2 (values for each

[4] We additionally tested the interaction between each predictor and time and found that the interaction between ba-da categoricity and time was significant, β = -0.91, SE = 0.32, z = -2.83, p = 0.005, suggesting that *more* categorical perception of the ba-da continuum predicted more accurate identification at both time points, but more strongly on the first day. Because the simple effects were not statistically significant and this pattern is not consistent with any theoretical predictions, we think this is likely a Type I error and advise caution in any interpretation of this finding.

predictor can be found in our supplementary analysis document on OSF, https://osf.io/2zg6c/), indicating no issues with multicollinearity. Finally, we acknowledge that we only included a very brief training period in this experiment design. Even though we saw significant improvement on the discrimination task after training and after the overnight interval, it is possible that predictors of real-world or longer-term learning of non-native speech contrasts differ from those found for short-term, laboratory-based studies. The results presented here cannot speak to possible relationships between native-language speech perception and long-term learning of non-native speech sounds.

Another possibility is that the relationship between categoricity and perception of unfamiliar speech sounds might be cue-specific or it might only hold for sounds that are perceptually similar in the native and non-native languages (for more on cue-specificity, see e.g., Kapnoula et al., 2021; Saito et al., 2022). This could at least explain why we did not find a significant relationship between s-sh categoricity and perception of the Hindi contrast. Though the sounds /b/ and /d/ are more similar to the Hindi sounds, some work has found that listeners show more categorical patterns of perception for stop consonants than other types of sounds (Eimas, 1963; Healy & Repp, 1982; Repp, 1981). Therefore, we may find a wider range of variability in categoricity of the s-sh continuum, especially when testing a larger sample. This could allow us to detect a relationship between categoricity and perception of a non-native contrast that is similar to the s-sh sounds.

Experiment 2

The purpose of Experiment 2 is twofold: First, to address the issue of statistical power raised in Experiment 1, we simplified the design and collected a much larger sample size. Second, we tested the question of whether the relationship between categoricity and discrimination of unfamiliar speech sounds is specific to perceptually similar contrasts. In this experiment, we tested participants' categoricity with the same phonetic continua (ba-da and s-sh) as in Experiment 1. We also tested participants' discrimination of the

Hindi dental and retroflex stop consonants and their discrimination of an additional unfamiliar speech contrast, namely the voiceless alveolo-palatal /ɕ/ and retroflex /ʂ/ fricatives found in Polish. Prior to explicit instruction, these speech sounds are not well discriminated by native English listeners (Lisker, 2001; McGuire, 2007; Żygis & Padgett, 2010), due to their similarity to the /ʃ/ category in English (Best et al., 2001; Kuhl et al., 2008). We therefore tested whether categoricity or consistency of native-language sounds predict discrimination of perceptually similar non-native sounds (i.e., whether perceptual measures of ba-da predicts discrimination of the Hindi sounds and s-sh predicts discrimination of the Polish sounds). We did not include a training task in this experiment, so this experiment focused on the relationship between categoricity/consistency and *naive* perception of non-native speech sound contrasts. We note, however, that the predictions from the perceptual assimilation model, for example, apply to naive perception (Best et al., 2001). In addition, in our previous work, we have seen that performance on a discrimination pretest typically predicts performance on the posttests after training (Fuhrmeister & Myers, 2020); therefore, it is possible that any findings from this experiment could be relevant for learning, but we did not test that explicitly here.

Method

Participants

A sample of 269 participants was recruited from the online data collection platform Prolific (www.prolific.co). Our goal was to have a sample size of at least 200 participants, and data collection was stopped after our sample size reached at least 200 after eliminating participants for the reasons described below. All participants included in data analysis were native speakers of North American English and had typical hearing. A total of 67 participants were excluded from data analysis for the following reasons: 32 for failing to use headphones during the experiment, 13 for learning a second language in which they were fluent before the age of 13 years, 6 for self-reported hearing difficulty, and 16 for

exposure to a language with a retroflex sound whether from family members, a caregiver, in a school environment, or through the media. These languages include unspecified Chinese ($n = 4$), Mandarin ($n = 3$), Vietnamese ($n = 3$), Hindi ($n = 2$), Russian ($n = 2$), Serbian ($n = 1$), and Sicilian ($n = 1$). The remaining sample was comprised of 202 participants. All participants gave informed consent according to the University of Connecticut Institutional Review Board and were paid $10 per hour for participation.

Stimuli

Perception of non-native speech sounds was assessed using Hindi and Polish phonemic contrasts. For the Hindi contrast, the same dental and retroflex stimuli from Experiment 1 were used (minimal pair non-words /d̪ug/ and /d̪̺ug/). Polish stimuli consisted of minimal pair, consonant-vowel syllables that differed in whether they began with the voiceless alveolo-palatal /ɕɑ/ or voiceless retroflex /ʂɑ/) consonant. Stimuli were recorded by a male, native speaker of Polish in a soundproof booth and were rescaled to a mean amplitude of 65dB sound pressure level in Praat (Boersma & Weenink, 2013). All tasks in Experiment 2 were presented to participants remotely on a laptop or desktop computer through Gorilla Experiment Builder (www.gorilla.sc, Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2018).

The stimuli used to measure the perception of native-language speech categories were the same seven-step fricative (sign-shine) and stop consonant (ba-da) continua from Experiment 1.

Procedure

Participants were routed from Prolific to Gorilla to complete the experiment remotely between the dates of July 4, 2021 and October 20, 2021. After providing consent, participants completed a headphone check in which they listened to three pure tones in stereo and were asked to indicate which tone was the quietest (Woods, Siegel, Traer, &

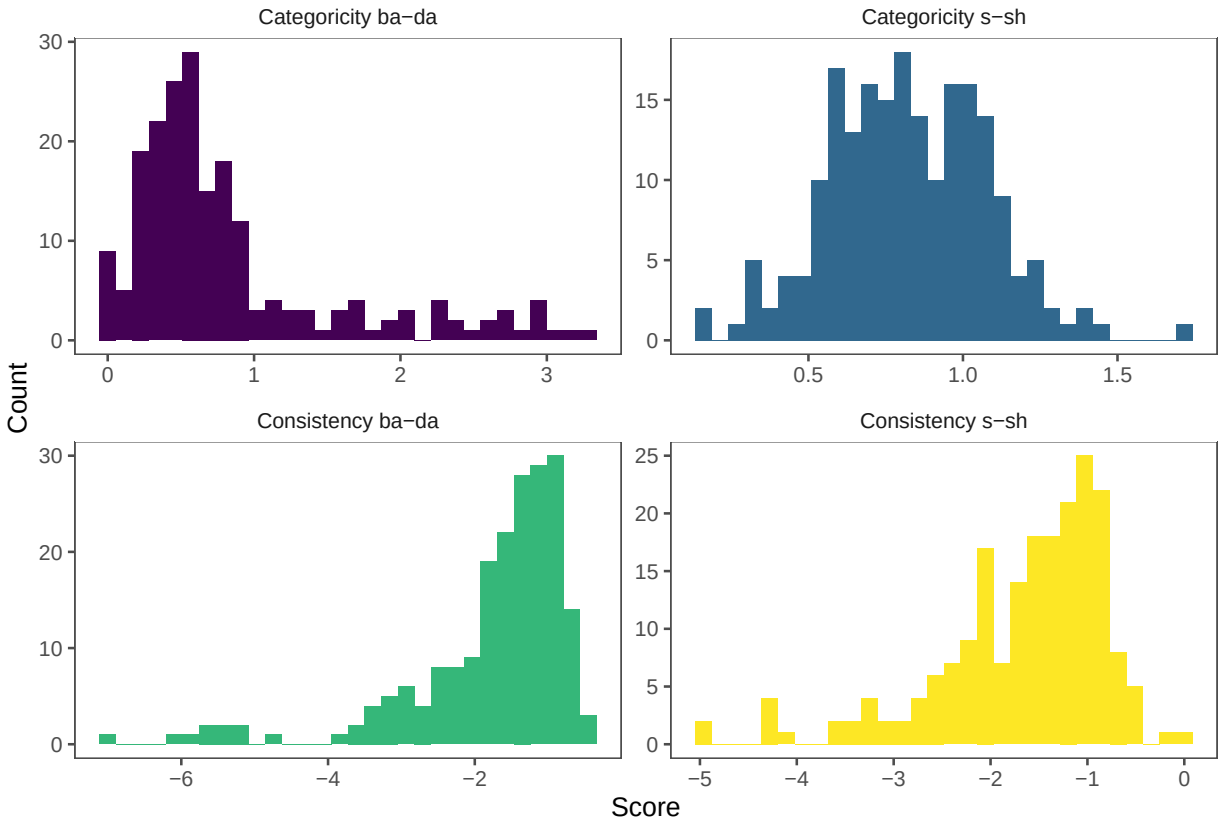

Figure 3. Histograms of categoricity and consistency measures on each continuum. Note the differences in axes among the plots.

McDermott, 2017). This task utilizes phase cancellation in the stimuli and can determine whether the participant is listening to the stimuli through headphones or over loudspeakers. If participants passed the headphone check they continued with the experiment, and if they failed, they were asked to redo the headphone check a second time after being reminded to wear headphones. If participants failed the headphone check the second time, they completed the experiment although their data was excluded from analysis. Participants then completed two questionnaires collecting information about demographics and language experience and completed the four phonetic tasks (Hindi AX discrimination, Polish AX discrimination, ba-da categorization, and sign-shine categorization). To address potential concerns that the order of tasks in Experiment 1 may have influenced performance on the visual analogue scale tasks (specifically that hearing the Hindi sounds perceived as unusual

exemplars of /d/ would make participants more gradient in their perception of the ba-da continuum), we set the order of tasks in this experiment so that no participant would hear similar native and non-native stimuli in two tasks that were completed consecutively. Specifically, participants completed the tasks in one of the following orders, to which they were randomly assigned: (1) Hindi AX, Polish AX, ba-da categorization, sign-shine categorization; (2) Polish AX, Hindi AX, sign-shine categorization, ba-da categorization; (3) ba-da categorization, sign-shine categorization, Hindi AX, Polish AX; (4) sign-shine categorization, ba-da categorization, Polish AX, Hindi AX. We opted for counterbalancing task order rather than having participants perform the native phonetic tasks first because we think it is equally possible that the task with the visual analogue scale would influence discrimination of the non-native contrasts. For example, we can imagine that drawing listeners' attention to subtle acoustic differences in sounds in their native language might provide them with a strategy for discriminating non-native sounds. Using a Bayesian *t*-test and Bayes factor test, we did not find evidence that categorization slopes differed on the ba-da continuum between participants depending on whether native or non-native perceptual tasks were completed first ($BF_{10} = 1.77$), and we found moderate evidence for *no* difference on the s-sh continuum ($BF_{10} = 0.17$, see for example Schönbrodt & Wagenmakers, 2018; Wetzels & Wagenmakers, 2012, for Bayes factor interpretations).

Non-native speech sound learning tasks. Two AX discrimination tasks just like the one described in the previous experiments were administered to test participants' discrimination accuracy of the Hindi and Polish sounds[5]. Each task consisted of 64 trials.

Native-language speech perception measures. To measure categoricity and consistency of native-language speech sounds, we administered a similar task to the visual

[5] The inter-stimulus interval for the discrimination task of the Polish sounds was 500 ms, which differed from the Hindi task. Although the interstimulus interval can influence performance on a discrimination task, the inter-stimulus intervals from both tasks were in the range of peak discrimination performance discussed in Gerrits and Schouten (2004).

analogue scale described in Experiment 1. Due to constraints from the online experiment presentation software, we did not have a slider for participants to move along a line; rather, participants saw a bar with numbers 1-7 (the number of points on the continuum) and the corresponding words/syllables on each side of the bar. They then clicked on a number from 1-7 to represent where between the two phonemes they thought the token belonged. This task was similar to the visual analogue scale in Experiment 1 in that participants saw the numbers 1-7 on a visual scale between the two phonemes.

Analysis approach

The slopes of the logistic curve to measure categoricity and the measure of consistency in each continuum were calculated as described in Experiment 1. Figure 3 shows the distributions of the categoricity and consistency measures. We again computed all possible correlations of these measures: categoricity between the two continua, $r = 0.32$, $p = <.001$; consistency between the two continua, $r = 0.34$, $p = <.001$; ba-da categoricity and consistency, $r = 0.29$, $p = <.001$; and s-sh categoricity and consistency, $r = 0.13$, $p = 0.07$.

We followed the same analysis procedure as in Experiment 1 to predict discrimination accuracy of the Hindi and Polish contrasts. We fit separate models for each contrast. For each model, the dependent variable was whether participants answered "different" (0 or 1), and fixed effects included a deviation coded contrast for condition (same = -.5, different = .5), the interaction of condition and categoricity (mean centered), and the interaction of condition and consistency (mean centered).

Table 3

Results of the Hindi discrimination task in Experiment 2.

Predictors	*Estimate*	*SE*	*z*	*p*
Intercept	-0.14	0.04	-3.43	<.001

Condition (d')	0.58	0.08	6.90	<.001
Condition x Categoricity	-0.02	0.07	-0.24	0.81
Condition x Consistency	0.10	0.05	2.15	0.03

Table 4

Results of the Polish discrimination task in Experiment 2.

Predictors	*Estimate*	*SE*	*z*	*p*
Intercept	-0.20	0.09	-2.24	0.03
Condition (d')	1.17	0.17	6.70	<.001
Condition x Categoricity	0.02	0.20	0.12	0.91
Condition x Consistency	0.30	0.06	5.19	<.001

Results

Results of the model predicting discrimination accuracy of the Hindi contrast are summarized in Table 3. Similar to the previous experiments, we see that the intercept is significantly different from zero with participants answering "same" more often, and there is a main effect of condition, suggesting participants answered "different" more often on different trials. More interestingly for the current purposes, we observe a significant interaction of condition and consistency, suggesting that more consistent responders on the ba-da task discriminated the Hindi sounds more accurately. The interaction of condition and categoricity did not reach significance. Variance inflation factors were all close to 1, suggesting no issues with multicollinearity.

Results of the model predicting discrimination accuracy of the Polish sounds are summarized in Table 4. We again see a significant intercept, main effect of condition, and interaction of condition and consistency. This suggests that participants who responded more consistently on the s-sh task also discriminated the Polish sounds more accurately.

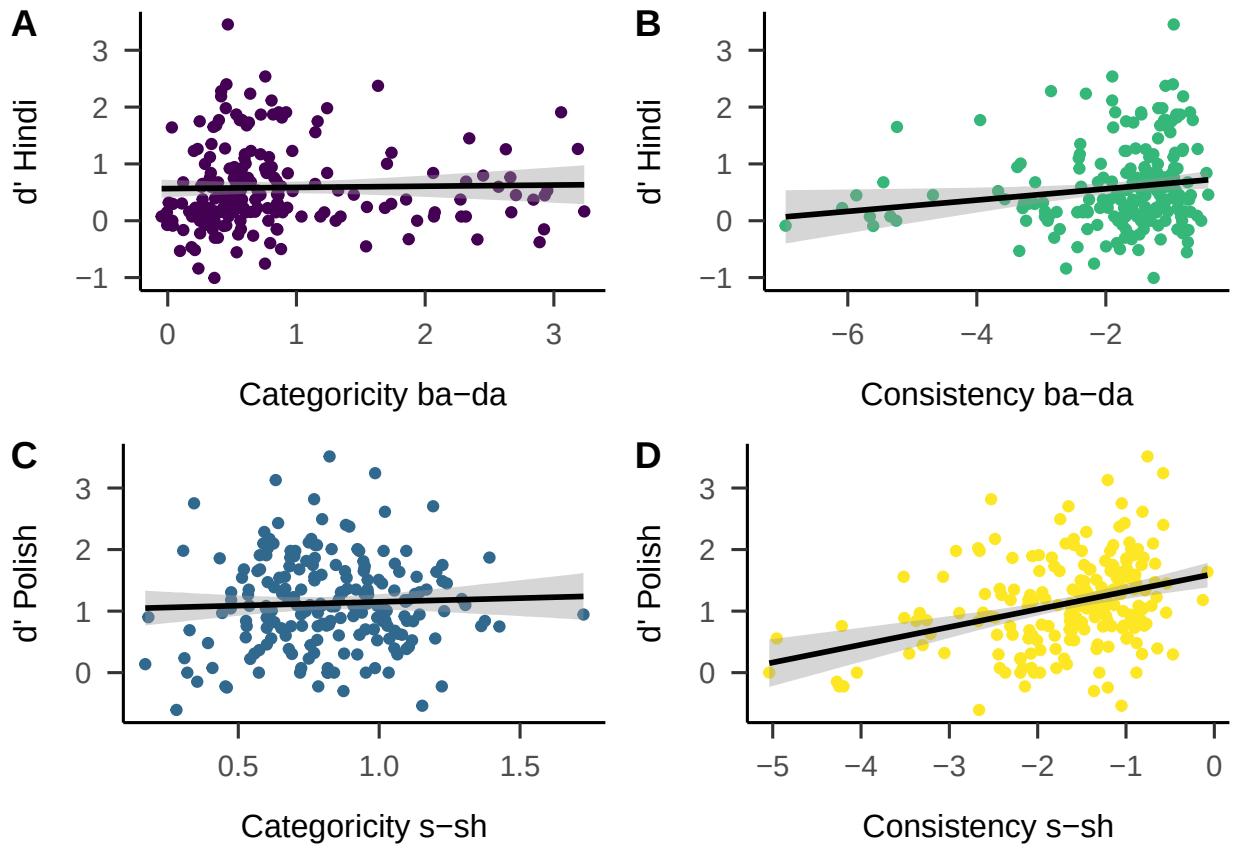

Figure 4. Relationships between discrimination accuracy of the Hindi contrast and A. ba-da categoricity and B. ba-da consistency. Relationships between discrimination accuracy of the Polish contrast and C. s-sh categoricity and D. s-sh consistency. We used d' scores for visualization purposes only; however, the statistical models predicted trial-level accuracy.

The interaction between condition and categoricity was not significant. Variance inflation factors were again all close to 1 in this model, suggesting no issues with multicollinearity. Results of Experiment 2 are displayed in Figure 4.

Discussion

Experiment 2 was a partial replication and extension of Experiment 1. With a much larger sample size than Experiment 1, we saw that consistency of responses on a native-language visual analogue scale task predicted better discrimination accuracy of a perceptually similar non-native speech contrast. This pattern of results was found for both

non-native contrasts we tested. This suggests that Experiment 1 was likely underpowered to detect this relationship. We return to this finding in the General Discussion.

Despite the larger sample in Experiment 2, categoricity still did not significantly predict discrimination accuracy of either non-native contrast. Once again, variance inflation factors for all predictors in each model were close to 1, suggesting the null result of categoricity is not simply due to multicollinearity. However, we cannot conclude there is no effect with a null result in the frequentist framework. For this reason, we did a small-scale Bayesian meta-analysis to synthesize the estimates in the two experiments (categoricity predicting naive discrimination accuracy in Experiment 1 and categoricity predicting discrimination accuracy for both sets of stimuli in Experiment 2; three estimates in total). Details on this analysis can be found in the appendix. To summarize briefly here, we performed Bayes factor tests and a sensitivity analysis (as recommended by several authors) to determine the relative evidence for the alternative hypothesis (an effect of categoricity on discrimination accuracy) over the null hypothesis under different assumptions (e.g., Hoijtink, Mulder, Lissa, & Gu, 2019; Kass & Raftery, 1995; Schad, Nicenboim, Bürkner, Betancourt, & Vasishth, 2021). We mostly found anecdotal evidence for the null hypothesis and only found moderate evidence for the null model when we use a very wide prior. Thus, we have evidence against a very large effect; however, if there is an effect of categoricity on non-native discrimination, it is possible that it is so small that we were still unable to find evidence for it even when synthesizing results across both experiments.

General discussion

Studies of non-native speech perception often find a wide range of individual differences in how accurately listeners perceive unfamiliar speech sounds. The goal of the current study was to test whether current theories of non-native speech perception can be extended to account for this variability in a particular way. We predicted that listeners

who show more graded perceptual patterns of speech sounds in their native language would discriminate difficult non-native speech sounds more accurately than those who show more categorical patterns of perception. In two different experiments, we did not find that individual differences predicted discrimination accuracy of non-native contrasts. Instead, we found that *consistency* of responses on a speech categorization task predicted discrimination accuracy.

Individual differences in categorical perception

In general, the current findings do not lend support to the idea that categoricity of native-language speech sounds is what makes differentiating non-native speech sounds hard at the *individual* level. On the basis of several theories of non-native speech sound learning (e.g., Best et al., 2001; Best & Tyler, 2007; Flege, 1995; Iverson et al., 2003; Kuhl, 1994; Kuhl et al., 2008), we reasoned that individual differences in categoricity would predict the degree to which individuals assimilate difficult non-native speech contrasts to a perceptually similar native-language speech sound. It is intuitive to assume that perceptual reorganization of speech in infancy (Kuhl et al., 2006; Werker & Tees, 1984) is related to the development of categorical perception. Specifically, if we lose perceptual sensitivity to certain non-native speech sounds and those non-native speech sounds are then assimilated to native-language speech categories during perception, it seems logical that categorical perception could be responsible for the assimilation of perceptually similar non-native sounds to native-language categories. However, categorical perception and perceptual assimilation may not be as related as is often assumed. This is consistent with the more modern view that speech perception follows a more protracted developmental pattern, in which perceptual representations continue to be refined (this includes becoming more graded) at least throughout adolescence (Hazan & Barrett, 2000; Idemaru & Holt, 2013; McMurray, 2022; McMurray et al., 2018; Nittrouer, 2004; Zevin, 2012). Thus, graded perception of speech seems to be the norm (McMurray, 2022; McMurray et al., 2018;

Toscano, McMurray, Dennhardt, & Luck, 2010), though individuals differ in the extent of this gradiency (Fuhrmeister & Myers, 2021; Kapnoula et al., 2021, 2017; Kong & Edwards, 2016). However, these differences may not have much of an impact on how accurately listeners can perceive non-native speech sounds.

It is possible, however, that the behavioral measures of categoricity that we used in the experiments presented here were not sensitive enough for measuring individual variability. For example, we used discrete points on the visual analogue scale, whereas previous studies have used continuous scales (Kapnoula et al., 2021, 2017; Kong & Edwards, 2016). Furthermore, in Experiment 2, we used a task more akin to a Likert scale, and we also found a correlation of 0.29 between the consistency and categoricity measures for the ba-da continuum. This could indicate that at least some participants were treating the task more like a two-alternative forced choice task, as consistency and categoricity measures are typically more highly correlated for two-alternative forced choice tasks (see e.g., Kapnoula et al., 2017; McMurray et al., 2018). Future work could use a more sensitive measure such as a visual analogue scale without discrete options or an online measure such as eye tracking to measure categoricity and test its relationship with non-native speech perception. Nonetheless, we did see a great deal of individual variability in the categoricity measure from the visual analogue scale that we used (see Figure 3), suggesting a decent amount of sensitivity for capturing individual differences. The distribution of scores does suggest that a handful of participants may have responded in a very categorical manner on the ba-da task; however, most participants did indeed show more graded patterns of perception, as the bulk of the distribution of scores falls between 0 and 1. It is also notable that we only saw this pattern with the stop continuum and not the fricative continuum, suggesting that this pattern of responses might have more to do with the nature of the sounds (e.g., Eimas, 1963; Healy & Repp, 1982; Repp, 1981) rather than the task itself.

Perceptual consistency

Results of the current study suggest that the most sensitive perceivers of novel non-native contrasts were those who had the most consistent categorization responses or ratings of native-language sounds. The fact that native-language response consistency was related to discrimination accuracy of non-native contrasts supports the hypothesis that trial-to-trial stability in the auditory response leads to stronger speech and language skills. This interpretation is consistent with findings in the field that general auditory and speech skills themselves support perception of both native and non-native speech sounds (and possibly also learning of new speech sounds). For example, Díaz, Baus, Escera, Costa, and Sebastián-Gallés (2008) reported electrophysiological differences in good vs. poor perceivers in response to both native and non-native speech sounds, which they interpreted as evidence of a speech-specific skill. Additionally, work from our lab and others has found that measures of phonological skills predict learning of a non-native speech sound contrast (Earle & Arthur, 2017; Fuhrmeister et al., 2020; see also Perrachione et al., 2011 for similar findings for learning of a non-native tonal contrast). More evidence for this idea comes from Heffner and Myers (2021), who found relationships between measures of non-native speech sound learning and adaptation to different types of speech in listeners' native language (rate and accent adaptation).

The relationship between consistency and discrimination accuracy might suggest that fidelity or stability of encoding of the stimulus is beneficial for accurate discrimination of non-native speech sounds. This parallels the findings from studies using the auditory brainstem response, which is a measure of how accurately the brainstem encodes an auditory stimulus (e.g., Skoe & Kraus, 2010). Fidelity of neural responses such as the auditory brainstem response has been linked to individual differences in speech perception abilities (Bidelman, Moreno, & Alain, 2013), non-native speech sound learning (Kachlicka et al., 2019; Omote, Jasmin, & Tierney, 2017), and individual differences in reading

abilities among unimpaired readers (Skoe, Brody, & Theodore, 2017). It is possible that we are seeing a similar phenomenon behaviorally: Listeners who encode familiar speech sounds more accurately or with higher fidelity may be able to leverage this same skill to distinguish subtle differences in unfamiliar sounds.

The findings presented here are also reminiscent of some findings of consistency in speech production. For example, several studies have shown that the precision or compactness of pronunciations of native-language vowels predicts perception (Kartushina & Frauenfelder, 2013) and production (Kartushina & Frauenfelder, 2014; Kartushina, Hervais-Adelman, Frauenfelder, & Golestani, 2016) accuracy of a non-native vowel contrast. In fact, Flege and Bohn (2021) recently published a revised version of the speech learning model, and based partially on these findings that production precision predicts learning of second-language speech categories, this version adds the "L1 category precision hypothesis" to the theory. This hypothesis asserts that the more precise the first-language categories are when a learner is exposed to a second language, the more likely the learner will be to establish new phonetic categories in the second language because the first and second-language sounds will be more discernible when the first-language categories are more compact/precise. Although the revised speech learning model defines category precision in terms of speech *production*, our findings suggest that consistency in *perception* may also be relevant.

According to the revised speech learning model, speech category precision may be related to auditory processing and acuity (Flege & Bohn, 2021). This is consistent with other studies that have found links between auditory abilities and non-native speech sound learning. For instance, Kidd, Watson, and Gygi (2007) found support for a general auditory ability for recognizing speech and non-speech sounds. Lengeris and Hazan (2010) observed that a non-speech frequency discrimination task correlated with perception of first- and second-language speech contrasts, which they take to argue that auditory

processing can explain individual differences in speech perception. Finally, Saito et al. (2022) found that more precise discrimination accuracy of acoustic cues that differentiate non-native speech sounds resulted in more accurate perception of those sounds.

The literature on the auditory brainstem response, speech production, and now our results from a speech perception task all seem to converge to suggest that consistency or precision of speech representation at multiple levels partially explains individual differences we observe in production and perception of non-native speech sounds. In sum, the category precision hypothesis in the revised speech learning model (and related accounts that prioritize the stability and precision of phonetic category representations) holds promise for the future of individual differences research in non-native speech sound learning.

Conclusions

The findings presented here do not support the assumption that categorical perception of native-language speech sounds is what is responsible for individual differences in perception of non-native speech sounds. In contrast, our findings do suggest at least some degree of overlap of native and non-native perceptual systems, specifically with respect to consistency or precision of representations.

References

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2018). Gorillas in our midst: Gorilla. Sc, a new web-based experiment builder. *bioRxiv*, 438242.

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

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

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, *109*(2), 775–794.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by english-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 345.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, *1334*, 1–47.

Bidelman, G. M., Moreno, S., & Alain, C. (2013). Tracing the emergence of categorical speech perception in the human auditory system. *Neuroimage*, *79*, 201–212.

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer [computer program]. Version 5.3. 51. *Online: Http://Www. Praat.org*, *2*.

Bowles, A. R., Chang, C. B., & Karuzis, V. P. (2016). Pitch ability as an aptitude for tone learning. *Language Learning*, *66*(4), 774–808.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training japanese listeners to identify english/r/and/l: Long-term retention of learning in

perception and production. *Perception & Psychophysics*, *61*(5), 977–985.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Burnham, D. K., Earnshaw, L. J., & Clark, J. E. (1991). Development of categorical identification of native and non-native bilabial stops: Infants, children and adults. *Journal of Child Language*, *18*(2), 231–260.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

Díaz, B., Baus, C., Escera, C., Costa, A., & Sebastián-Gallés, N. (2008). Brain potentials to native phoneme discrimination reveal the origin of individual differences in learning the sounds of a second language. *Proceedings of the National Academy of Sciences*, *105*(42), 16083–16088.

Drouin, J. R., Theodore, R. M., & Myers, E. B. (2016). Lexically guided perceptual tuning of internal phonetic category structure. *The Journal of the Acoustical Society of America*, *140*(4), EL307–EL313.

Earle, F. S., & Arthur, D. T. (2017). Native phonological processing abilities predict post-consolidation nonnative contrast learning in adults. *The Journal of the Acoustical Society of America*, *142*(6), EL525–EL531.

Eimas, P. D. (1963). The relation between identification and discrimination along speech and non-speech continua. *Language and Speech*, *6*(4), 206–217.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, *92*, 233–277.

Flege, J. E., & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). *Second Language Speech Learning: Theoretical and Empirical Progress*, 3–83.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language*, *41*(1), 78–104.

740 Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (Third). Thousand
741 Oaks CA: Sage.
742 Fuhrmeister, P., & Myers, E. B. (2020). Desirable and undesirable difficulties: Influences of
743 variability, training schedule, and aptitude on nonnative phonetic learning. *Attention,*
744 *Perception, & Psychophysics*, *82*(4), 2049–2065.
745 Fuhrmeister, P., & Myers, E. B. (2021). Structural neural correlates of individual
746 differences in categorical perception. *Brain and Language*, *215*, 104919.
747 Fuhrmeister, P., Phillips, M. C., McCoach, D. B., & Myers, E. B. (2022 November 252022
748 November 25). Relationships between native and non-native speech perception.
749 Retrieved from https://osf.io/2zg6c/
750 Fuhrmeister, P., Schlemmer, B., & Myers, E. B. (2020). Adults show initial advantages
751 over children learning difficult non-native speech sounds. *Journal of Speech, Language,*
752 *and Hearing Research*, *63*(8), 2667–2679.
753 Gerrits, E., & Schouten, M. E. (2004). Categorical perception depends on the
754 discrimination task. *Perception & Psychophysics*, *66*(3), 363–376.
755 Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies:
756 Some arguments on why and a primer on how. *Social and Personality Psychology*
757 *Compass*, *10*(10), 535–549.
758 Golestani, N., Molko, N., Dehaene, S., LeBihan, D., & Pallier, C. (2007). Brain structure
759 predicts the learning of foreign speech sounds. *Cerebral Cortex*, *17*(3), 575–582.
760 Golestani, N., Paus, T., & Zatorre, R. J. (2002). Anatomical correlates of learning novel
761 speech sounds. *Neuron*, *35*(5), 997–1010.
762 Golestani, N., & Zatorre, R. J. (2009). Individual differences in the acquisition of second
763 language phonology. *Brain and Language*, *109*(2-3), 55–67.
764 Hattori, K., & Iverson, P. (2009). English/r/-/l/category assimilation by japanese adults:
765 Individual differences and the link to identification accuracy. *The Journal of the*
766 *Acoustical Society of America*, *125*(1), 469–479.

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics*, *28*(4), 377–396.

Healy, A. F., & Repp, B. H. (1982). Context independence and phonetic mediation in categorical perception. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(1), 68.

Heffner, C. C., & Myers, E. B. (2021). Individual differences in phonetic plasticity across native and nonnative contexts. *Journal of Speech, Language, and Hearing Research*, *64*(10), 3720–3733.

Hoijtink, H., Mulder, J., Lissa, C. van, & Gu, X. (2019). A tutorial on testing hypotheses using the bayes factor. *Psychological Methods*, *24*(5), 539.

Idemaru, K., & Holt, L. L. (2013). The developmental trajectory of children's perception and production of english/r/-/l. *The Journal of the Acoustical Society of America*, *133*(6), 4232–4246.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Kettermann, A., Siebert, C., et al. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57.

Kachlicka, M., Saito, K., & Tierney, A. (2019). Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound. *Brain and Language*, *192*, 15–24.

Kapnoula, E. C., Edwards, J., & McMurray, B. (2021). Gradient activation of speech categories facilitates listeners' recovery from lexical garden paths, but not perception of speech-in-noise. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(4), 578.

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(9), 1594.

Kartushina, N., & Frauenfelder, U. H. (2013). On the role of L1 speech production in L2 perception: Evidence from spanish learners of french. *Proceedings of the 14th Interspeech Conference*, 2118–2122.

Kartushina, N., & Frauenfelder, U. H. (2014). On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in Psychology*, *5*, 1246.

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2016). Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics*, *57*, 21–39.

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

Kempe, V., Bublitz, D., & Brooks, P. J. (2015). Musical ability and non-native speech-sound processing are linked through sensitivity to pitch and spectral information. *British Journal of Psychology*, *106*(2), 349–366.

Kidd, G. R., Watson, C. S., & Gygi, B. (2007). Individual differences in auditory abilities. *The Journal of the Acoustical Society of America*, *122*(1), 418–435.

Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, *59*, 40–57.

Kuhl, P. K. (1994). Learning and representation in speech and language. *Current Opinion in Neurobiology*, *4*(6), 812–822.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2), F13–F21.

Lengeris, A., & Hazan, V. (2010). The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of english vowels for native speakers of greek. *The Journal of the Acoustical Society of America*, *128*(6), 3757–3768.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358.

Lim, S., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science*, *35*(7), 1390–1405.

Lisker, L. (2001). Hearing the polish sibilants [s š ś]: Phonetic and auditory judgements. *Travaux Du Cercle Linguistique de Copenhague XXXI. To Honour Eli Fischer-Jørgensen*, 226–238.

Luthra, S., Fuhrmeister, P., Molfese, P. J., Guediche, S., Blumstein, S. E., & Myers, E. B. (2019). Brain-behavior relationships in incidental learning of non-native phonetic categories. *Brain and Language*, *198*, 104692.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324.

Mayr, R., & Escudero, P. (2010). Explaining individual variation in L2 perception: Rounded vowels in english learners of german. *Bilingualism: Language and Cognition*, *13*(3), 279–297.

McGuire, G. (2007). English listeners' perception of polish alveopalatal and retroflex voiceless sibilants: A pilot study. *UC Berkeley PhonLab Annual Report*, *3*(3).

McMurray, B. (2022). *The acquisition of speech categories: Beyond perceptual narrowing,*

beyond unsupervised learning and beyond infancy.

McMurray, B., Danelz, A., Rigler, H., & Seedorff, M. (2018). Speech categorization develops slowly through adolescence. *Developmental Psychology*, *54*(8), 1472.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42.

Miller, J. L. (1997). Internal structure of phonetic categories. *Language and Cognitive Processes*, *12*(5-6), 865–870.

Nittrouer, S. (2004). The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. *The Journal of the Acoustical Society of America*, *115*(4), 1777–1790.

Omote, A., Jasmin, K., & Tierney, A. (2017). Successful non-native speech perception is linked to frequency following response phase consistency. *Cortex*, *93*, 146–154.

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1), 461–472.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2021). *nlme: Linear and nonlinear mixed effects models*. Retrieved from https://CRAN.R-project.org/package=nlme

Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, *29*(2), 191–215.

Polka, L. (1991). Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions. *The Journal of the Acoustical Society of America*, *89*(6), 2961–2977.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Razzaghi, M. (2013). The probit link function in generalized linear models for data mining

875 applications. *Journal of Modern Applied Statistical Methods*, *12*(1), 19.

876 Repp, B. H. (1981). Two strategies in fricative discrimination. *Perception &*

877 *Psychophysics*, *30*(3), 217–227.

878 Saito, K., Kachlicka, M., Suzukida, Y., Petrova, K., Lee, B. J., & Tierney, A. (2022).

879 Auditory precision hypothesis-L2: Dimension-specific relationships between auditory

880 processing and second language segmental learning. *Cognition*, *229*, 105236.

881 Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2021).

882 Workflow techniques for the robust use of bayes factors. *arXiv Preprint*

883 *arXiv:2103.08744*.

884 Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a

885 priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*,

886 *110*, 104038.

887 Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue

888 use in production and perception of a non-native sound contrast. *Journal of Phonetics*,

889 *52*, 183–204.

890 Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning

891 for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142.

892 Silbert, N. H., Smith, B. K., Jackson, S. R., Campbell, S. G., Hughes, M. M., & Tare, M.

893 (2015). Non-native phonemic discrimination, phonological short term memory, and

894 word learning. *Journal of Phonetics*, *50*, 99–119.

895 Skoe, E., Brody, L., & Theodore, R. M. (2017). Reading ability reflects individual

896 differences in auditory brainstem function, even into adulthood. *Brain and Language*,

897 *164*, 25–31.

898 Skoe, E., & Kraus, N. (2010). Auditory brainstem response to complex sounds: A tutorial.

899 *Ear and Hearing*, *31*(3), 302.

900 Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency:

901 Does musical ability matter? *Psychological Science*, *17*(8), 675–681.

Stevens, K. N., & Blumstein, S. E. (1975). Quantal aspects of consonant production and perception: A study of retroflex stop consonants. *Journal of Phonetics*, *3*(4), 215–233.

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, *10*(4), 277–303.

Thompson, C. G., Kim, R. S., Aloe, A. M., & Becker, B. J. (2017). Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic and Applied Social Psychology*, *39*(2), 81–90.

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, *21*(10), 1532–1540.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63.

Wetzels, R., & Wagenmakers, E.-J. (2012). A default bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064.

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.

Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods*, *41*(2), 257–267.

Zevin, J. D. (2012). A sensitive period for shibboleths: The long tail and changing goals of speech perception over the course of development. *Developmental Psychobiology*, *54*(6), 632–642.

Żygis, M., & Padgett, J. (2010). A perceptual study of polish fricatives, and its implications for historical sound change. *Journal of Phonetics*, *38*(2), 207–226.

Appendix

Meta-analysis of estimates in the two experiments

In two different experiments, we did not see that categoricity significantly predicted discrimination accuracy of a non-native speech contrast. However, with a null result in the frequentist framework, we cannot claim to have *no* effect. Furthermore, we saw a fair amount of between-experiment variability, as evidenced by the range in standard errors of the effect of interest. We therefore conducted a Bayesian meta-analysis of the effect (interaction between condition and categoricity) that included the experiments presented in current study. Meta-analyses are useful for synthesizing results of several experiments, and they estimate the effect of interest and the precision of that effect using the estimates and standard errors from the individual studies. We acknowledge that we have a limited number of data sets for a meta-analysis; however, Goh, Hall, and Rosenthal (2016) advocate specifically for what they call a "mini meta-analysis," in which authors perform a meta-analysis on similar experiments reported in a paper. We further chose to perform the meta-analysis in the Bayesian framework in order to perform Bayes factor tests to test whether we have relative evidence for the alternative hypothesis (an effect of categoricity on discrimination accuracy) over the null hypothesis.

Data sets and extraction of estimates

Meta-analyses synthesize data across several experiments using the estimate of the effect in question (here the interaction of categoricity and condition) and its standard error. We included three estimates and their standard errors in this meta-analysis. For Experiment 1, we fit a model using only the data from the discrimination pretest because that is common to both experiments. We additionally only used the measure of categoricity from the ba-da continuum from Experiment 1 to be more like Experiment 2. For each data set (Experiment 1, Experiment 2 Hindi contrast, Experiment 2 Polish contrast), we fit a generalized linear mixed effects model with a probit link that predicted

whether the participant answered "different" (0 = no, 1 = yes). Fixed effects included condition (deviation coded, different = .5, same = -.5) and the interaction of condition and categoricity (mean centered). When possible, the full random effects structure was included and random effects were removed one by one in case the model did not converge without warnings. The estimates and standard errors from the interaction effect were extracted from each model and entered into the meta-analysis.

Meta-analysis

We conducted a random effects meta-analysis (as opposed to a fixed effects meta-analysis), which assumes a different true effect for each data set (Sutton & Abrams, 2001). To do this, we fit a Bayesian generalized linear mixed effects model that estimates an intercept from the estimates of the effects from each data set weighted by their standard error with a random intercept for data set. This model was fit using the brms package (Bürkner, 2017). To test whether we have evidence for this effect (the interaction of categoricity and condition), we did a Bayes factor test using bridge sampling, which informs us on the relative evidence for one model over another (in this case the alternative vs. a null model). Bayes factors are supposed to be taken as graded evidence: A Bayes factor of 1 indicates no evidence for one model over another, Bayes factors between 1 and 3 or between 1/3 and 1 are considered anecdotal or inconclusive evidence for one model over another, and Bayes factors over 3 or less than 1/3 are typically considered to indicate at least moderate evidence for one model over the other (Schönbrodt & Wagenmakers, 2018; Wetzels & Wagenmakers, 2012).

The Bayesian framework allows us to incorporate prior beliefs into the statistical model. Bayes factors are sensitive to the priors used in the model, so we did a sensitivity analysis, in which we report Bayes factors for a range of priors, as recommended by several authors (e.g., Hoijtink et al., 2019; Kass & Raftery, 1995; Schad et al., 2021). The effect of interest here is the interaction effect (interaction of categoricity and condition), and for

that effect, we chose a normal prior with a mean of 0 and a range of standard deviations from .05 to .5. For the standard deviation, we chose regularizing priors from a normal distribution with a mean of 0 and standard deviation of 1.

Table A1

Results of the meta-analysis and sensitivity analyses. We report the prior, estimate, credible interval (CrI), and bayes factor in favor of the alternative hypothesis (BF10).

Prior	*Estimate*	*CrI*	*BF10*
Normal (0,0.05)	0.005	[-0.09, 0.09]	0.90
Normal (0,0.1)	0.011	[-0.15, 0.16]	0.76
Normal (0,0.2)	0.017	[-0.24, 0.26]	0.53
Normal (0,0.3)	0.017	[-0.31, 0.33]	0.40
Normal (0,0.4)	0.015	[-0.37, 0.36]	0.32
Normal (0,0.5)	0.013	[-0.4, 0.39]	0.26

Results and discussion

Results of the meta-analysis including sensitivity analyses are summarized in Table A1. We report the prior for the effect of interest used in each Bayes factor test, the posterior mean (i.e., meta-analytic estimate), the 95% credible interval, and the Bayes factor in favor of the alternative hypothesis.

Most of the evidence is inconclusive; however, when we assume a larger effect size (prior width of 0.5), we see moderate evidence for the null hypothesis. In no case did we see evidence for the alternative hypothesis. At the very least, we can conclude that we have evidence against a large effect. It is nonetheless possible that there is a true *small* effect of categoricity on discrimination accuracy of a non-native contrast; however, we may need even more data than we have collected in the two experiments here to obtain evidence for

993 such a small effect.