Exploring Individual Differences in Native Phonetic Perception and Their Link to Non-Native Phonetic Perception

Claire T. Honda[1,2], Meghan Clayards[2,3,4], and Shari R. Baum[2,3]

[1]Integrated Program in Neuroscience, McGill University

[2]Centre for Research on Brain, Language and Music, Montreal, Canada

[3]School of Communication Sciences and Disorders, McGill University

[4]Department of Linguistics, McGill University

Author Note

All data from these experiments are publicly available on the Open Science Framework (OSF) and can be accessed at https://osf.io/ez5qh/?view_only=8e4a1498e04f4ee0946752ee93b9ce71. The hypotheses, methods, and analyses for Experiment 2 were preregistered and are available at the same link.

The authors have no conflict of interest to declare.

This work was supported by a grant awarded by the Natural Sciences and Engineering Research Council of Canada (NSERC) to Shari R. Baum, a grant awarded by the Social Sciences and Humanities Research Council (SSHRC) to Meghan Clayards, and NSERC Canada Graduate Scholarship-Master's (CGS M) and Postgraduate Scholarship-Doctoral (PGS D) grants along with a McGill Faculty of Medicine Max Binz Fellowship awarded to Claire T. Honda.

Correspondence concerning this article should be addressed to Claire T. Honda, Integrated Program in Neuroscience, McGill University, 2001 McGill College Avenue, 8th floor, Montreal, QC, H3A 1G1, Canada. Email: claire.honda@mail.mcgill.ca.

CRediT Authorship Contribution Statement

Claire T. Honda: Conceptualization, Methodology, Investigation, Formal analysis, Visualization, Writing – original draft. Meghan Clayards: Conceptualization, Methodology, Supervision, Writing – review & editing. Shari R. Baum: Conceptualization, Methodology, Supervision, Writing – review & editing.

Abstract

Adults differ considerably in their perception of both native and non-native phonemes. For instance, when presented with continua of native phonemes on 2-alternative forced choice (2AFC) or visual analogue scaling (VAS) tasks, some people show sudden changes in responses (i.e., steep identification slopes) and others show gradual changes (i.e., shallow identification slopes). Moreover, some adults are more successful than others at learning unfamiliar phonemes. The predictors of these individual differences and the relationships between them are poorly understood. It also remains unclear to what extent different tasks (2AFC vs. VAS) may reflect distinct individual differences in perception. In two experiments, we addressed these questions by examining the relationships between individual differences in performance on native and non-native phonetic perception tasks. We found that shallow 2AFC identification slopes were not related to shallow VAS identification slopes but were related to inconsistent VAS responses. Additionally, our results suggest that consistent native perception may play a role in promoting successful non-native perception. These findings help characterize the nature of individual differences in phonetic perception and contribute to our understanding of how to measure such differences. This work also has implications for encouraging successful acquisition of new languages in adulthood.

Public Significance Statement: Successfully perceiving speech sounds is a crucial skill for spoken communication; yet individuals show differences in how they perceive both native and non-native speech sounds. We studied the relationships between performance on different native and non-native speech perception tasks, finding that (a) different tasks measure different subtleties and (b) people with consistent perception of native speech sounds tend to be better at

accurately perceiving non-native sounds. These findings have implications for understanding the nature of individual differences in speech perception and for helping adults to learn new languages successfully.

Keywords: phonetic perception, non-native perception, individual differences, gradient perception, consistency

It is well established that there are individual differences in speech perception. Even healthy young adults show differences in how they perceive speech in their native language. For example, some people have better perception of speech in noise compared to others (Surprenant & Watson, 2001). Similarly, some people show greater perceptual plasticity, i.e., an increased ability to successfully adapt their perception to changes in speaking rate or accent (Heffner & Myers, 2021). People also differ in the extent to which their speech perception is affected by different factors, such as coarticulation (Yu & Lee, 2014) or visual information about the speaker's mouth movement (as in the McGurk effect; Strand et al., 2014). Differences even in phonetic perception—an elemental building block of higher-level speech perception—have been documented for decades, such as differences in the categorization of stops (Hazan & Rosen, 1991) and in the discrimination of sibilants (Perkell et al., 2004).

While it is interesting to note these differences, current research is attempting to understand what underlies them, and in doing so to better understand speech perception. For example, some researchers have proposed that differences in basic auditory processing play a role (Cumming et al., 2015; Won et al., 2016). Others have found links between differences in executive function and in speech perception (Kapnoula et al., 2017; Kim et al., 2020). Such

studies help us to understand the broader architecture of speech perception. One goal of the current paper is to better understand sources of individual differences as measured by two speech perception tasks – two-alternative forced choice (2AFC) and visual analog scaling (VAS) – by comparing different measures of performance on each one across a large sample of participants. In doing so, we test the hypothesis that the tasks each reflect distinct individual differences in speech sound perception.

In addition to the individual differences in native phonetic perception described above, there are differences in non-native phonetic perception. Adult learners of non-native phoneme contrasts show great variability in performance, with some successfully distinguishing contrasts and others having great difficulty even after receiving training and feedback (e.g., Bradlow et al., 1997; Hanulíková et al., 2012; Hattori & Iverson, 2010; Strange & Dittman, 1984). Non-native perception has been shown to depend in part on numerous factors, including native language background (e.g., Flege et al., 1997), musical ability (e.g., Slevc & Miyake, 2006), and auditory acuity measures such as temporal processing (Kempe et al., 2012) or formant and pitch discrimination (Kachlicka et al., 2019); however, the impact of these factors seems to depend on the particular non-native sounds being perceived, and accounts for only a portion of the variation in performance. As such, the predictors of successful non-native phonetic perception remain relatively poorly understood. The second goal of this paper is to test whether differences in native speech sound perception predict discrimination of difficult second language sound contrasts encountered for the first time.

Individual Differences in Native Speech Perception

One of the most ubiquitous methods for measuring phonetic perception is using 2-alternative forced choice (2AFC) tasks. In these tasks, participants are generally presented with a

continuum of speech stimuli (e.g., ranging in small steps from *bet* to *bat*) and must classify each stimulus into one category or the other. When a participant's average response to each stimulus is plotted against actual changes in stimulus properties, the result yields an identification slope that can range from shallow to steep (indicating that responses are changing gradually/sharply with changes in stimuli).

There are differences between people in terms of how shallow or steep their identification slopes are. Shallower slopes on 2AFC tasks have previously been linked to various language impairments (Manis et al., 1997; Joanisse et al., 2000; Serniclaes et al., 2001; Werker & Tees, 1987) and to illiteracy (Serniclaes et al., 2005), and have accordingly been considered to reflect an unsuccessful and undesirable pattern of perception compared to steeper slopes. Shallow slopes have been thought to reflect poorly defined boundaries between phonemic categories, potentially due to enhanced discrimination within categories (Serniclaes et al., 2001), whereas steep slopes have been thought to reflect sharply defined boundaries between categories. The association between shallow slopes and language impairment has therefore led to the suggestion that sensitivity to within-category, sub-phonemic detail can be maladaptive. However, it is not clear whether shallow 2AFC slopes actually reflect fine-grained, within-category sensitivity. Instead, they might reflect an inconsistent ability to perceive or categorize sounds (Kapnoula et al., 2017; Serniclaes et al., 2001). Thus, it may be erroneous to relate within-category sensitivity to impairment (see Kapnoula et al., 2017 and Apfelbaum et al., 2022 for other examples of this point). In contrast to shallow slopes, steep slopes on 2AFC tasks are often assumed to indicate categorical perception, which has been proposed as an effective solution to the problem of how continuous cues in the acoustic signal are mapped onto discrete categories during perception (Liberman et al., 1957).

Limits of a Categorical View of Perception

Empirically, categorical perception refers to the observation that (1) when presented with a continuum that ranges in equal steps from one category to another, people tend to perceive a sharp distinction between categories (i.e., a steep identification slope as described above); and that (2) stimuli belonging to the same category are often discriminated more poorly than equivalently distant stimuli that cross a category boundary (Liberman et al., 1957). This finding has led to the theoretical view that our perceptual representations are warped based on our top-down knowledge of categories, facilitating processing (Goldstone & Hendrickson, 2010).

Despite the fact that the theory of categorical perception has been hugely popular and influential, there is also widespread evidence challenging it (see McMurray, 2022 for a review). Even from the early days of its proposal, categorical perception was not observed for all speech sounds (Fry et al., 1962) and there was evidence that task demands were at least partly responsible for the phenomenon (Pisoni & Tash, 1974; Hary & Massaro, 1982). Since then, work using behavioural, eye-tracking, and neurophysiological techniques has led to a growing consensus that auditory encoding and speech perception are in fact inherently gradient (McMurray et al., 2008; McMurray et al., 2002; Miller, 1994; Ou & Yu, 2022; Toscano et al., 2010). Gradient perception refers to an ability to distinguish gradual, fine-tuned phonetic differences rather than sudden phonemic ones as in categorical perception. For example, using category goodness ratings, Miller (1994) demonstrated that phonetic categories have a gradient and context-dependent structure; some stimuli are perceived as better exemplars of a category than others, and this perception can flexibly change when relevant contextual factors (e.g., speech rate, syllable structure, or lexical status) are altered. In a similar vein, McMurray and colleagues found that identification slopes became less steep (more gradient) when words were

used instead of meaningless syllables, when pictures were used instead of letters, and when four alternatives were used instead of two (McMurray et al., 2008). Thus, the common finding of steep (or "categorical") slopes on 2AFC tasks being associated with successful perception may stem largely from task demands; after all, the task requires a categorical response, so it is natural for it to elicit categorical-looking response patterns in successful listeners. Given all of this evidence, it is useful to note that *categorization of speech sounds* is a necessary process to derive meaning from the speech signal and does not preclude gradient perception, while *categorical perception* (as a theoretical view involving perceptual warping) is not necessary to explain the patterns of responses that have been observed on tasks such as 2AFC.

A Less Categorical Measurement: The VAS Task

Unlike 2AFC tasks which elicit a categorical decision, visual analogue scaling (VAS) tasks require the participant to indicate what they heard along a continuous line between two options (Massaro & Cohen, 1983). VAS tasks provide a valuable alternative to 2AFC tasks for a variety of reasons. For instance, VAS tasks appear to have superior psychometric properties to 2AFC tasks. Munson et al. (2017) found that fricative ratings on a 2AFC task differed in the extent to which they were influenced by particular acoustic cues, depending on whether the 2AFC ratings were interleaved with more continuous ratings (gender typicality of speech) or more categorical ratings (which category the adjacent vowel belonged to, among 5 options); in contrast, ratings of the same stimuli on a VAS task did not differ depending on these biasing conditions. VAS ratings therefore seem to be less influenced by concurrent tasks (Munson et al., 2017).

Critically, VAS tasks may be better suited to studying the phenomenon of gradient vs. categorical perception; they enable responses that are closely related to the acoustic

characteristics of speech (Apfelbaum et al., 2022; Massaro & Cohen, 1983; Munson et al., 2012) and that correlate with continuous measures of production (Schellinger et al., 2017). Using a VAS task, Kong & Edwards (2016) found clear differences between participants' response patterns (some participants had more gradient-looking responses, others were more categorical), showing the task's potential as an alternative to the 2AFC format for studying individual differences in phonetic perception.

Relationships Between Individual Differences in 2AFC and VAS

Although they have been much studied, 2AFC slopes on their own are not very informative for reasons described below. However, by comparing data from both 2AFC and VAS tasks, it is possible to better understand the nature of the individual differences underlying different response patterns on these two tasks. Kapnoula et al. (2017) did precisely this, comparing participants' identification slopes on 2AFC and VAS tasks. They found that the slopes on the two tasks were not related within participants, suggesting that the tasks do not measure the same construct (Kapnoula et al., 2017). Note that they used different ways of estimating slopes for the two tasks, an issue we will return to later. Furthermore, Kapnoula et al. (2017) measured how consistently participants responded on the VAS task by calculating the difference between a given participant's actual response on each trial and their predicted response based on their VAS identification slope, and then calculating the standard deviation of these residuals for each participant. Interestingly, they found that shallower 2AFC slopes were marginally related to less consistent VAS responses. Their interpretation was that a shallow 2AFC slope may reflect inconsistent perception of speech sounds rather than actual gradiency of perception (Kapnoula et al., 2017).

To illustrate these findings and the limitations of 2AFC slopes, consider a listener with very gradient perception—that is, with fine-tuned sensitivity to within-category differences between sounds. When presented with a 2AFC task, such a listener might show very different identification slopes depending on their response strategy. One strategy would be to categorize the sounds consistently based on whichever response option they more closely resemble, which would result in a steep identification slope (sharp distinction between categories). Another strategy would be to respond probabilistically by matching the proportion of their two responses to the degree that the sound matches the two alternatives, which would result in a shallow identification slope (Clayards et al., 2008). On the 2AFC task, two very different identification slopes can thus arise from the same underlying perception of speech sounds. Furthermore, a shallow slope on the 2AFC task could arise due to two possibilities: the participant could have more signal-driven, gradient perception and be responding probabilistically as just described, or they could have more category-driven perception but be responding in a noisy and inconsistent way. These possibilities cannot be disambiguated without additional information from another task.

Now consider how the same listener with more gradient perception would respond on a VAS task. Unlike for the 2AFC task, there would be no ambiguity; the listener would show a shallow identification slope. Similarly, the VAS task can distinguish between whether the listener's perception is truly gradient—evidenced by a shallow slope—or in fact inconsistent—evidenced by dissimilar ratings for the same stimulus across trials. By comparing participants' slopes and consistency across the two tasks, it is therefore possible to determine whether 2AFC slopes reflect gradiency or consistency of perception, and whether a given participant's perception is more gradient or more categorical. In finding that 2AFC slopes were weakly related

to VAS consistency but not to VAS slopes, the work by Kapnoula et al. (2017) provides preliminary evidence that 2AFC slopes may tap more into the construct of consistency whereas VAS slopes tap more into the construct of gradiency.

The relationship between shallow 2AFC slopes and inconsistency of perception provides a potential explanation for why the previously mentioned studies have linked shallow 2AFC slopes to language impairment. Thus, it is potentially problematic to use the term *gradient* when referring to shallow 2AFC slopes or to associate the concept of gradient/less-categorical perception with impairment (e.g., Manis et al., 1997; Werker & Tees, 1987) when the true issue may lie in inconsistent perception. For this reason, we will refer to identification slopes as being shallow or steep—terms that do not assume a direct association between slope and the construct of gradiency/categoricity—rather than gradient or categorical. Because these terms have unbiased interpretations and facilitate comparisons of results across tasks, we will often use them to refer to slopes derived both from 2AFC and from VAS tasks. This being said, VAS tasks naturally allow for a continuous/gradient form of responding that is more likely to reflect true gradiency compared to 2AFC responses (Apfelbaum et al., 2022), so we will occasionally follow previous work in referring to measures of gradiency when such measures have been derived from VAS tasks. Note, however, that some authors use gradiency to also refer to shallow 2AFC or 4AFC slopes (e.g., Ou et al., 2021; Ou & Yu, 2022).

The Nature and Potential Functions of Gradiency

Gradiency (as measured by VAS tasks) appears to be a relatively consistent property of the individual. It has been shown to be related across different testing sessions using the same stimuli (Kong & Edwards, 2016), across different contrasts (Fuhrmeister & Myers, 2021; Kapnoula & McMurray, 2021; but see Kapnoula et al., 2021 for contrasting evidence), and

across native and non-native perception (Kong & Kang, 2022). Individual differences in gradiency may reflect anatomical differences in auditory processing architecture, since they relate to differences in cortical surface area (Fuhrmeister & Myers, 2021) and in how cues are neurally encoded and transformed along the auditory pathway (Kapnoula & McMurray, 2021; Ou & Yu, 2022).

Interestingly, various lines of evidence point to the idea that gradiency may not be an indicator of unsuccessful perception as previously thought. For instance, gradiency can reflect experience-related sensitivity to fine acoustic detail, with trained speech-language pathologists giving VAS ratings that are more closely related to acoustic characteristics of the signal compared to inexperienced listeners (Munson et al., 2012). In addition, more gradient VAS responses have been associated with an increased ability to integrate multiple acoustic cues in the speech signal (Kapnoula et al., 2017; Kapnoula & McMurray, 2021; Kim et al., 2020; Kong & Edwards, 2016; Kong & Kang, 2022). Gradiency thus relates to the ability to integrate multiple acoustic cues and to perceive fine-tuned changes in those cues, which appears to encourage perceptual flexibility in the face of ambiguous input (Clayards et al., 2008; Desmeules-Trudel & Zamuner, 2019).

In line with the notion that gradiency promotes perceptual flexibility, Kapnoula et al. (2021) found that listeners with shallower VAS slopes showed greater recovery from lexical garden paths during an eye-tracking task. For example, when presented with a stimulus such as *pumpernickel* in which the initial consonant had been manipulated to sound ambiguous between [p] and [b], such listeners were more likely to switch their gaze from a competitor item (*bumpercar*) to the appropriate target item compared to listeners with steeper VAS slopes (Kapnoula et al., 2021). In other words, by being sensitive to fine-grained acoustic details, the

gradient listeners were more readily able to reconsider and flexibly adjust their initial interpretation of misleading stimuli. Further support comes from work that has demonstrated a relationship between inhibitory control and gradiency (Kapnoula et al., 2021). Greater inhibitory control appears to promote gradiency by enabling listeners to manage ambiguous input that activates competing phonemic representations, thus granting listeners greater perceptual flexibility (Kapnoula et al., 2021). The flexibility afforded by gradiency could have a range of benefits given that flexible perception is useful for adapting to variation in both native and non-native speech (Heffner & Myers, 2021)—successful listeners must constantly adapt to differences in the speech signal that arise from numerous factors such as speaking rate, coarticulation, speaker gender, and accent.

Individual Differences in Non-Native Speech Perception

As discussed above, adult learners of non-native phoneme contrasts show great variability in performance, and this variation is not fully accounted for by the factors that have been identified so far. At early learning stages, learners often start out with vastly different scores on tests of non-native perceptual ability; and even those with similar baseline scores often go on to show very different outcomes after non-native perceptual training (e.g., Bradlow et al., 1997; Golestani & Zatorre, 2009; Hanulíková et al., 2012).

Differences in native phonetic perception are one potential predictor of non-native perception. Individuals with better discrimination of native vowels have been shown to have better identification of non-native vowels on a ten-alternative forced-choice task (Lengeris & Hazan, 2010). Similarly, greater sensitivity to native contrasts on a gating task has been related to better discrimination of non-native Mandarin tones (Kalaivanan et al., 2023). Other work suggests that having clearly defined, compact representations of a native vowel in

psychoacoustic space predicts greater sensitivity to a non-native vowel contrast (Kogan & Mora, 2022). There is also recent neurophysiological evidence that sensitivity to native contrasts is positively correlated with sensitivity to non-native contrasts (Norrman et al., 2022).

It is not surprising, then, that existing models of non-native phonetic learning emphasize the influence of native phonetic categories. The perceptual assimilation model (Best & Tyler, 2007), speech learning model (Flege, 1995), native language magnet model (Kuhl et al., 2008), and perceptual interference model (Iverson et al., 2003) all describe how a learner's difficulty with a given non-native phoneme will depend on the similarity between that phoneme and native phonemes. For example, one prediction that has received some support is that the difference between two non-native speech sounds is easier to distinguish when the non-native sounds are perceptually assimilated to two different native categories, compared to when they are assimilated to the same native category (Best & Tyler, 2007; Mayr & Escudero, 2010). These models address which phonemes are easier or harder for learners of a given language background overall, without directly addressing individual differences in success between learners. However, some studies have used these models as a framework to predict the success of non-native perception based on differences in assimilation patterns. Mayr and Escudero (2010) studied how native English speakers assimilated German vowels to native categories. They found variety in assimilation patterns, with some participants perceiving the German contrasts in terms of a single native category and others perceiving them in terms of two or more native categories. Importantly, these differences in assimilation were predictive of identification success: participants who assimilated the German contrasts to two distinct native categories showed better identification of those contrasts than participants who assimilated them to a single native category (Mayr & Escudero, 2010). Hattori and Iverson (2009) similarly observed individual

differences in assimilation patterns for native Japanese speakers perceiving the English /ɹ/-/l/ contrast. While they did not find a relationship between assimilation patterns and identification success for the English contrast, they did find that identification success was predicted by differences in participants' representations of the third formant for /ɹ/ and /l/. The aforementioned models can thus provide some insight into links between native and non-native perception at the individual level. Moreover, by relating native categories to non-native sound learning, the models imply that differences in non-native perception should be predicted not only by assimilation patterns, but also by differences in the perception of native categories.

As an example, more gradient responses to native sounds on VAS tasks could indicate less of an influence of language-specific categories on perception, and thus yield an easier time learning new categories. Furthermore, gradiency may reflect fine-tuned and flexible perception as detailed above, which could conceivably assist with the discrimination of non-native phonemes. Conversely, steeper identification slopes on 2AFC tasks might predict better non-native perception, since an optimal strategy for a gradient listener on such tasks could be to clearly label each sound based on whichever category it best fits (as discussed above). Fuhrmeister et al. (2023) recently studied non-native discrimination ability and native gradiency as measured by a VAS task and were surprised not to find evidence for a relationship between the two. Instead they found that non-native discrimination related to the consistency of VAS responses, i.e., how similar participants' ratings were across trials for a given stimulus. However, they used a VAS task resembling a Likert scale, with only 7 discrete points (in contrast to the continuous scales used by other researchers such as Kapnoula et al., 2017 and Kong & Edwards, 2016). The presentation of discrete response options may have incited participants to treat the task more similarly to a 2AFC task, putting into question whether the task was truly measuring

gradiency. Furthermore, Fuhrmeister et al. (2023) tested only consonants (no native or non-native vowels). Even though gradiency appears to be a relatively stable individual property that holds across different speech sounds as described above (Fuhrmeister & Myers, 2021; Kapnoula & McMurray, 2021; Kong & Edwards, 2016; Kong & Kang, 2022), certain sounds such as consonants are likely to elicit gradient responses to a lesser degree because listeners typically show greater sensitivity to within-category differences in vowels than in consonants (e.g., Fry et al., 1962; Schouten & Van Hessen, 1992). Perhaps a relationship did not emerge between native gradiency and non-native discrimination in their study because there was not a wide enough range of gradiency values due to the use of consonants alone, or not enough variability within the gradiency values due to the limited sensitivity of a 7-point scale. The relationships between native 2AFC and VAS performance and non-native discrimination thus remain to be clarified.

The Current Study

The current study had two primary aims. First, we wanted to clarify which individual differences are reflected in performance on 2AFC and VAS tasks. It is of interest to determine whether these two tasks measure the same construct—2AFC tasks are ubiquitous in psycholinguistic research, so it is important to understand what they may be tapping into and how they compare to other tasks. Some authors have concluded that 2AFC and VAS tasks do not measure the same construct, and that 2AFC responses relate to consistency rather than gradiency (e.g., Kapnoula et al., 2017). On the other hand, some authors have used the term gradiency when referring to 2AFC (Ou & Yu, 2022) or 4AFC (Ou et al., 2021) slopes, for example positing that such “gradiency” is in part due to how strongly one’s subcortical and cortical representations of speech correlate with one another (Ou & Yu, 2022); this assumes that 2AFC slopes do measure the same construct as VAS slopes. Furthermore, more gradient responses on a VAS task

have been related to greater use of secondary cues on a 2AFC task (Kapnoula et al., 2017; Kim et al., 2020). Steeper categorization of primary cues on a 2AFC task has additionally been linked to greater use of secondary cues on the same task (Clayards, 2018), and steeper slopes on a 4AFC task have similarly been linked to greater use of secondary cues in an eye-tracking task (Ou et al., 2021). Together, these findings seem to imply that steeper slopes on 2AFC tasks could be related to shallower slopes on VAS tasks. Such an inverse relationship might also be anticipated given that a gradient listener could show a steep 2AFC slope based on their response strategy, as outlined earlier. Developmental work by McMurray et al. (2018) has found that steeper identification functions and more gradient phonetic perception (as measured by eye-tracking) appear to develop in tandem during adolescence, further hinting at the possibility of an inverse relationship between 2AFC slopes and VAS slopes. However, it is also possible that the slopes are not related across the tasks if 2AFC responses relate more to inconsistency than gradiency, as tentatively proposed by Kapnoula et al. (2017).

Previous work that compared performance across the two tasks did not use identical continua of stimuli, instead presenting participants with VAS continua consisting of 35 stimuli and 2AFC continua consisting of only 14 stimuli (Kapnoula et al., 2017). This difference in the richness of continua across tasks could have contributed to the lack of relationship reported between 2AFC and VAS slopes; in order to more directly compare performance across the two tasks, exactly the same continua should be used for both. Kapnoula et al. (2017) also only found a marginal relationship between 2AFC slopes and VAS consistency (Kapnoula et al., 2017); a conceptual replication is needed in order to clarify whether this finding seems to be a spurious or a genuine one. Furthermore, different analysis techniques have been used across different tasks and across different studies, so it is unclear whether the results depend on the analysis techniques

(more on this in the *Comparing Slope Estimate Methods* section). The relationship between tasks and the individual differences reflected by each task therefore requires further investigation, bringing us to our first hypothesis.

Hypothesis 1: 2AFC and VAS tasks provide different ways of measuring individual differences in speech sound perception, with VAS slopes reflecting gradiency and 2AFC slopes reflecting consistency. If this is the case, 2AFC slopes will not relate to VAS slopes but will relate to the consistency of VAS responses, with inconsistent VAS performance predicting shallower 2AFC slopes.

The second question we aimed to address was whether discrimination of difficult non-native contrasts could be predicted by differences in native phonetic perception as measured by VAS and 2AFC tasks. We predicted that shallower VAS slopes and steeper 2AFC slopes might both reflect the ability to make accurate and fine-tuned judgments about acoustic cues and might therefore relate to better non-native discrimination. If shallow VAS slopes do predict better non-native perception abilities, this would support the notion that gradiency, as measured by VAS tasks, may actually be adaptive and beneficial. This brings us to our second hypothesis.

Hypothesis 2: The ability to discriminate finely tuned differences in native speech sounds relates to the ability to accurately distinguish non-native speech sounds. If this is the case, steeper 2AFC slopes and shallower VAS slopes will relate to better non-native phonetic perception.

These hypotheses were tested in two experiments. In both experiments, we measured how English-speaking participants responded to identical continua of native speech sounds when the sounds were presented in a 2AFC and a VAS task. This enabled a direct comparison of responses across tasks. We also evaluated the participants' ability to discriminate unfamiliar non-native

(German) phonemes to investigate potential predictors of good non-native perception. Finally, we collected measures of working memory and attention in order to account for variation in non-linguistic cognitive abilities. Other studies of native and non-native perception have not accounted for such factors (e.g., Fuhrmeister et al., 2023), and yet it is relevant to do so given that executive function has been found to modulate the gradiency of native perception (Kapnoula et al., 2017; Kapnoula & McMurray, 2021) and the success of second language learning outcomes (e.g., Kwakkel et al., 2021; Lee, 2016). These cognitive factors are also important to consider in light of prior evidence that the working memory demands of a task can affect participants' responses and thus bias the conclusions that we draw (Gerrits & Schouten, 2004).

Because we collected a large number of measures and because there were many possible comparisons and analysis techniques available, we treated the first experiment as exploratory. This allowed us to explore the data and to develop an analysis approach after data collection. The methods, exclusion criteria, and analyses established in Experiment 1 were then preregistered on the Open Science Framework (OSF; https://doi.org/10.17605/OSF.IO/9DKGQ) as Experiment 2. Experiment 2 allowed us to then test our hypotheses with a priori analysis decisions and a larger sample size, strengthening our conclusions. We also performed additional non-preregistered analyses on the data from both experiments that we had not considered in the preregistration. The Methods section below describes the preregistered analyses first, which consisted of canonical correlation and multivariate multiple regression to assess Hypothesis 1 and of multiple regression to assess Hypothesis 2. The non-preregistered analyses, outlined at the end of the methods, included additional canonical correlations and a principal component analysis.

Methods

Aside from the sample size of participants recruited, the methods for Experiment 1 and Experiment 2 were identical. Data for Experiment 1 were collected from September to November 2020 and data for Experiment 2 were collected in July 2021.

Participants

Participants were right-handed, aged 18-35, born and living in the United States or Canada, and had no history of head injury or of literacy, language, cognitive, or hearing impairments. All were monolingual speakers of English. Participants received monetary compensation ($12.50 USD) and signed an informed consent form. The entire study had a duration of approximately 1.25 hrs including breaks. The research protocol was approved by the Institutional Review Board of the Faculty of Medicine and Health Sciences of McGill University. All participants were recruited through the online platform Prolific.co and were required to have access to a computer to complete the study. While monolingual English speakers with computer access are unlikely to be a universally representative sample, such constraints were necessary to control for prior language experience and to present the experiment in a consistent way across participants. Given that a wider demographic range can be obtained when recruiting from online platforms such as Prolific compared to when recruiting university students, we believe that our results are relatively generalizable.

Experiment 1 Sample Size

56 participants (21 females) were recruited through the platform Prolific.co.

Experiment 2 Sample Size

Experiment 2 was designed to replicate the results of Experiment 1 using a larger sample size. An appropriate sample size was estimated through a triangulation of approaches. As an initial step, we reviewed the sample size in comparable studies, most notably that of Kapnoula et

al. (2017), which is closest to the current study and included a sample of 120 participants, leading to some marginal and some significant effects. As a second approach, we relied on Harrell's (2015) rule of thumb applied to our design, which includes 6 predictors; multiplying the 6 predictors by 15 participants per predictor yields a sample size estimate of 90 participants minimum. Finally, we computed a power analysis based on multiple regression with 6 predictors, which reflects our regression models testing Hypothesis 1 and Hypothesis 2 (in fact Hypothesis 1 involved multivariate multiple regression with more than one response variable, but each model within the multivariate model had a single response variable and 6 predictors as in the power analysis, and conducting such an analysis on a multivariate design would be overly complex). This power analysis (with power = 0.95, alpha = 0.05, and number of predictors = 6) using the power.f2.test function from the pwr package (Champely, 2020) in R revealed that a sample size of 120 is required in order to reliably detect effects of a similar size ($r \geq 0.4$) to those reported in related studies (e.g., Clayards, 2018; Grimaldi et al., 2014; Kong & Edwards, 2016). Therefore, we settled on a sample size of 120. In order to arrive at a final sample size of around 120 after taking into account participant exclusion based on language experience and data quality issues, we recruited 139 participants (97 females) through the platform Prolific.co.

Questionnaires

Information about demographics, language history and proficiency, and musical experience was collected through a questionnaire adapted from the *Language History Questionnaire* (LHQ 2.0; Li et al., 2013) and the *Montreal Music History Questionnaire* (MMHQ; Coffey et al., 2011).

Tasks

Participants completed five tasks: two measuring native phonetic perception, one measuring non-native perception, one measuring sustained attention, and one measuring working memory (all described further below). Participants completed these tasks online at home using the Gorilla Experiment Builder (www.gorilla.sc; Anwyl-Irvine et al., 2020), with their own headphones. In order to standardize online sound presentation and ensure an acceptable listening environment, participants completed a headphone screening before the other tasks (Woods et al., 2017).

Native Phonetic Perception Tasks

Participants completed two native phonetic perception tasks. The tasks involved listening to minimal pairs that varied in different phonological contrasts (*bet-bat* and *dear-tear*; stimuli from Clayards 2018, publicly available at https://osf.io/369my/). These two pairs were selected because they enabled us to test perception of both a vowel and a consonant contrast, and they have successfully been used in the past to study individual differences in phonetic perception (Clayards, 2018). The minimal pairs were manipulated so that each one varied systematically in two acoustic cues relevant to the contrast (formant frequency and vowel duration for *bet-bat*, voice onset time and onset fundamental frequency for *dear-tear*). Each cue varied in 5 steps, and each version of the first cue was paired with each version of the second cue, leading to 25 stimuli per pair. This results in some ambiguous and some clear stimuli (stimuli whose cue values are both at the extremes—i.e., step 1 or step 5—sound clear and unambiguous; stimuli with more intermediate cue values sound more ambiguous). Details of stimulus properties are listed in Table 1, and further details of stimulus construction can be found in Clayards (2018). The same stimuli were used in both native phonetic perception tasks.

In the 2AFC task, participants indicated via mouse click which of two words they heard on each trial (e.g., *bet* or *bat*; side of the screen counterbalanced across participants). In the VAS task, participants were shown a slider on the computer screen with a word at each end (ends of the slider counterbalanced across participants). Participants indicated where along the continuous scale they perceived the stimulus to be (values were coded from 0 at one end to 100 at the other end, but were not displayed to participants during the task). Each stimulus from each minimal pair was presented 5 times in each task, for a total of 250 stimuli per task. Stimuli were blocked so that all 25 stimuli per pair appeared in a random order before any stimulus was repeated. *Bet-bat* and *dear-tear* trials were mixed in each block. All participants completed the VAS task first to avoid biasing responses based on the more categorical demands of the 2AFC task.

Non-native Phonetic Perception Task

In the non-native perception task, participants differentiated German vowels and consonants (ø: vs. œ, y: vs. ʏ, ʃ vs. ç in the International Phonetic Alphabet) which are known to be perceptually challenging speech sounds for native English speakers (Mayr & Escudero, 2010). German words containing these phonemes were presented in a 3-interval oddity (3-I oddity) task, in which participants heard three stimuli in a row and indicated which one (if any) was different. 3-I oddity tasks are useful for studying non-native phonetic perception since they do not require the participant to explicitly know the nature of the differences between unfamiliar stimuli (Strange & Shafer, 2008). They also have an advantage over similar tasks such as AXB, in that they are more intuitive for participants and their level of chance performance is lower (25% instead of 50%), allowing for greater variability in scores (Grimaldi et al., 2014). The complete set of German minimal pairs used in the task is found in Table 2.

In order to construct the stimuli for the task, three native German speakers were recorded producing each German word 5 times. The 1st and 5th productions were then discarded to leave 3 productions of each word per speaker. Sound files were edited to leave 20 ms before and after each production, and maximum amplitudes were normalized across speakers using GoldWave version 6.15 (GoldWave Inc., 2015). Each trial contained three words, one from each speaker, with an interstimulus interval of 500 ms. Participants indicated which word sounded different by clicking "1", "2", or "3" on a computer screen, or clicking "None" if all three words sounded the same. Half of the trials were switch trials where one of the words was the other member of the minimal pair, and the other half were catch trials where all three words were the same. For example, for the minimal pair "selig" and "seelisch" (/ˈzeːlɪç/ and /ˈzeːlɪʃ/), participants might hear "selig, seelisch, selig" on a switch trial and "selig, selig, selig" on a catch trial. There were 12 trials (6 switch and 6 catch trials) per minimal pair and 14 minimal pairs, for a total of 168 trials. Speaker order, odd speaker out, and odd minimal pair member were balanced across trials, and trial order was randomized. Before implementing the task, piloting with 6 participants was conducted in order to check for floor or ceiling effects. Piloting revealed overall accuracy rates of 39-65% (keeping in mind that chance performance is 25%), falling within the range of previous studies (e.g., Rauber et al., 2005; Silveira, 2011).

Table 1

Stimulus properties for the native perception tasks

bet-bat			**dear-tear**	
Formant frequencies of spectral steps (Hz)		Duration steps (ms)	Voice onset time steps (ms)	Onset F0 steps (Hz)
F1	F2			
625	1677	100	10	185
647	1610	140	20	195
663	1560	180	30	205
682	1546	220	40	215
740	1556	260	50	225

Table 2

German minimal pairs used in the 3-I oddity task

Consonant contrast		Vowel contrast 1		Vowel contrast 2	
Palatal fricative (ç)	Postalveolar fricative (ʃ)	Tense high front rounded vowel (yː)	Lax high front rounded vowel (ʏ)	Tense mid front rounded vowel (øː)	Lax mid front rounded vowel (œ)
Fichte /fɪçtə/	fischte /fɪʃtə/	Brühl /bʁyːl/	brüll /bʁʏl/	blöke /bløːkə/	Blöcke /blœkə/
Kirche /kɪəçə/	Kirsche /kɪəʃə/	Düne /dyːnə/	dünne /dʏnə/	gewöhne /gəvøːnə/	gewönne /gəvœnə/
Löchern /løçɪən/	löschern /løʃɪən/	fühlen /fyːlən/	füllen /fʏlən/	Höhle /høːlə/	Hölle /hœlə/
selig /zelɪç/	seelisch /zelɪʃ/	Hüte /hyːtə/	Hütte /hʏtə/	Söhne /zøːnə/	Sönne /zœnə/
Wicht /vɪçt/	wischt /vɪʃt/	Wüste /vyːstə/	wüsste /vʏstə/		

Cognitive Tasks

Finally, participants completed a version of the Continuous Performance Test (CPT; Conners et al., 2003) and a working memory task in order to assess whether any observed relationships between performance on the other tasks might be driven by individual differences in non-linguistic cognitive factors rather than in perception.

In the AX-CPT, participants were presented with a string of letters. They had to press a particular key whenever they saw the letter X preceded by the letter A (this was the case for 70% of trials) and press a different key in any other case (with keys counterbalanced across participants). There were 140 AX trials (A followed by X), 20 AY trials (A followed by a consonant other than X), 20 BX trials (B followed by X), and 20 BY trials (B followed by a consonant other than X), for a total of 200 trials.

A backwards digit span task was used to assess working memory (Wechsler, 2008). In this task, participants heard recorded series of numbers (presented with a 1s interstimulus interval) and were then asked to type them out in the reverse order. The number of digits to be recalled increased every 3 trials, starting with 2 digits and increasing to a maximum of 10 digits.

The task was terminated whenever the participant incorrectly answered all 3 trials of a given difficulty level.

Reliability of Measures

Given our focus on individual differences, one important consideration is whether the measures being used here are reliable within participants. To address this, we calculated the split-half reliability of each of our measures, adjusted with the Spearman-Brown correction. These reliability values are displayed in Supplemental Table S.1, revealing good reliability of all measures apart from the VAS slopes (this is simply due to bad fitting when not enough data is provided; see Supplemental Materials for further details).

Test-retest reliability is another informative measure of reliability. Common measures of test-retest reliability include Cronbach's alpha, test-retest correlation coefficients, and intraclass correlations which exist in ten different forms depending on the data structure and the type of reliability being calculated (Koo & Li, 2016). It is important to interpret reliability values according to the particular research context, and so we refer to benchmarks from the field of psychology: Cronbach's alpha and test-retest correlation values of 0.7-0.79, 0.80-0.89, and > 0.90 indicate fair, good, and excellent reliability respectively; and intraclass correlation values of 0.4-0.59, 0.6-0.74, and > 0.75 indicate fair, good, and excellent reliability respectively (Cicchetti, 1994). Previous studies have found fair to excellent test-retest reliability for various perceptual measures related to the ones used here, including auditory discrimination (Christopherson & Humes, 1992: Cronbach's alpha = 0.795; Saito & Tierney, 2022: intraclass correlation = 0.625; Wang & Humes, 2008: test-retest correlations > 0.90), sensitivity to the McGurk effect (Strand et al., 2014: test-retest correlation = 0.77), magnitude of the Ganong effect (Giovannone & Theodore, 2023: test-retest correlation = 0.72), use of a VAS scale (Brietzke et al., 2021:

intraclass correlation = 0.50), consonant identification (Geller et al., 2021: intraclass correlation = 0.80), and weighting of acoustic cues (Idemaru et al., 2012: test-retest correlation = 0.69; Souza et al., 2018: no difference in cue weightings across two sessions, as determined by Wilcoxon signed-ranks analyses). Importantly, individuals' 2AFC identification slopes for stimuli varying in voice onset time (VOT) and fundamental frequency (F0)—two of the same acoustic cues varying in our stimuli—have shown good to excellent reliability across sessions (Schertz et al., 2015: test-retest correlations = 0.90 for VOT and 0.84 for F0), and individuals' gradiency of speech perception on a VAS task has shown fair to good reliability across ratings of the same stimulus (Munson et al., 2021: intraclass correlation > 0.5 for 89% of listeners, average = 0.73). Furthermore, fair to good test-retest reliability has been shown for the backwards digit span task (Fox-Fuller et al., 2022: intraclass correlation = 0.66; Müller et al., 2012: intraclass correlation = 0.64; Wechsler, 2008: r = 0.71; Woods et al., 2011: r = 0.81), while fair to excellent test-retest reliability has been observed for the AX-CPT task (Barch et al., 2009: intraclass correlation = 0.81, Cooper et al., 2017: intraclass correlation = 0.70; Halperin et al., 1991: test-retest correlations = 0.65-0.74; Kraus et al., 2020: intraclass correlation = 0.72).

Based on both split-half and test-retest reliability, we can therefore conclude that our measures are reliable and appropriate for use in the context of individual differences studies such as the present one.

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. All stimuli, tasks, questionnaires, program code, and analysis methods developed by others have been cited in-text and included in the References section. The research materials (tasks and questionnaires) described above are available upon request. Stimuli from the

native phonetic perception tasks are publicly available, on the OSF page (https://osf.io/369my/) for Clayards (2018). The raw data for both experiments, along with the code needed to process and analyze it, is publicly available on the OSF (https://osf.io/ez5qh/?view_only=8e4a1498e04f4ee0946752ee93b9ce71). The design, hypotheses, and analysis plan of Experiment 2 were preregistered based on Experiment 1 and are available on the same OSF page.

Analysis and Results

Here we include tables and figures displaying results of primary interest. Additional tables and figures (for example, of model validation) can be found in the Supplemental Materials and in the R Markdown document on the OSF page for this project.

Data Exclusion

Participants who reported having phonetic training or being exposed to German were excluded (two participants in Experiment 1, 19 participants in Experiment 2), as this could affect performance on the non-native perception task. Participants were also excluded on a task-by-task basis depending on performance-based criteria. Criteria are outlined in the OSF preregistration (https://doi.org/10.17605/OSF.IO/9DKGQ). The total number of participants included in a given analysis is reported at the bottom of the figure (in the case of canonical correlation) or table (in the case of regression) displaying the output of that analysis.

Preparatory Data Analysis

Before conducting primary analyses, various preliminary analyses were carried out to obtain variables of interest.

Native Phonetic Perception Tasks

Slopes from the 2AFC task were calculated by fitting two mixed-effects logistic regression models to participant responses (one for bet/bat, one for dear/tear). Responses were coded as 0 for bet/dear and 1 for bat/tear. The fixed effects for each model were the first acoustic cue (which varied in 5 steps) and the second acoustic cue (which also varied in 5 steps) for the contrast in question, both of which were coded as continuous numeric variables and centered. The grouping factor was participant. The following correlated random effects were included in each model: by-participant random intercepts, and by-participant random slopes for the first acoustic cue and the second acoustic cue. The by-participant random slopes coefficients for each acoustic cue were extracted as the four variables of interest, since they quantify how much each participant differs from the group average (i.e., from the fixed effect coefficient) in their use of a given cue when categorizing stimuli (Clayards, 2018; Kong & Edwards, 2015). Larger random slopes coefficients (steeper slopes) for a given cue indicate greater use of that cue when categorizing stimuli. This analysis was carried out in R (R Development Core Team, 2020), using the lme4 package (Bates et al., 2015). The R syntax for the models described above was: glmer(2AFC response ~ Acoustic cue 1 step + Acoustic cue 2 step + (Acoustic cue 1 step + Acoustic cue 2 step | Participant), family = "binomial", control = glmerControl(optimizer = "bobyqa")).

Slopes from the VAS task were calculated by fitting the rotated logistic developed by Kapnoula et al. (2017) to participants' responses. The rotated logistic is conceptually similar to the 2AFC logistic regression coefficients mentioned above, but it models gradiency independently of acoustic cue use (since our stimuli vary in two acoustic cues). It is based on a four-parameter logistic function with estimates for minimum and maximum asymptotes, slope, and crossover point, but with one additional parameter: θ, which represents the angle of the

crossover point. The coordinate space is rotated to be orthogonal to this angle, with the result that the slope parameter provides a single measure of gradiency which is independent of the two acoustic cues constituting the space. These analyses were conducted in MatLab (version 2015a, The MathWorks Inc., USA). For each participant and each minimal pair, the average of the 5 responses to each of the 25 different stimuli in the VAS task was calculated, and the equation for the rotated logistic was fit to these averages. This resulted in two slope measures per participant: one for bet-bat responses, and one for dear-tear responses. Larger slope values from the rotated logistic function reflect shallower slopes and therefore more gradient responses.

To calculate differences in the consistency of participants' acoustic cue encoding, the rotated logistic was fitted to each participant's unaveraged responses. For each trial, the difference between the participant's actual VAS response and the response predicted by the rotated logistic was calculated. The standard deviation of these residuals was then averaged per minimal pair to provide two estimates of consistency per participant: one for bet-bat responses, and one for dear-tear responses. Greater standard deviation of residuals reflects less consistent responses. This is the same method used by Kapnoula et al. (2017) to calculate consistency, and closely resembles the method used by Fuhrmeister et al. (2023) who also calculated residuals from a logistic function fit to participants' VAS responses (but theirs was a regular rather than a rotated logistic function, since their continua varied only along one acoustic dimension).

Non-Native Perception Task

To quantify differences in non-native phonetic perception, the non-parametric sensitivity index *A* (a corrected version of *A'*; Zhang & Mueller, 2005) was calculated across performance on the fricative contrast and the vowel contrasts from the 3-I oddity task. This score is based on hits (correctly selecting the odd item in a switch trial) and false alarms (incorrectly selecting an

odd item in a catch trial). An *A* score of 1.0 indicates perfect discrimination, while a score of 0.5 indicates null discrimination. The calculation was done by implementing Zhang & Mueller's (2005) equation in R.

Cognitive Tasks

As a measure of sustained attention, a bin score was calculated from each participant's AX-CPT responses (Hughes et al., 2014). Unlike traditional reaction time (RT) difference measures, bin scores take into account both RT and accuracy, making them more reliable and suitable for use in individual differences studies (Draheim et al., 2019).

In preparation for bin scoring, trials were labeled by type (AX/AY/BX/BY) and also labeled as nonswitch (AX) or switch (AY/BX/BY). Only rows corresponding to the second letter of each trial were kept (i.e., X/Y, not A/B), and reaction times (RTs) were cleaned: for each participant, RTs < 200 ms were replaced with that participant's mean RT value, and RTs > 3 SD above their mean RT were replaced with a cutoff value of 3 SD above the mean. To calculate a participant's bin score, their mean RT on non-switch (AX) trials was subtracted from their RT for each switch trial (AY/BX/BY trials). The resulting RT differences were placed into ten bins which were assigned values ranging from 1 (smallest RT differences) to 10 (largest RT differences). Inaccurate responses were placed in a "bad" bin with a value of 20 to provide a penalty for low accuracy. Finally, the bin values for all of the participant's trials were summed to produce a final bin score. Lower bin scores indicate better attention due to smaller RT differences and/or higher accuracy.

From the backwards digit span test, the highest number of digits successfully recalled was taken as a measure of working memory.

Descriptive Overview

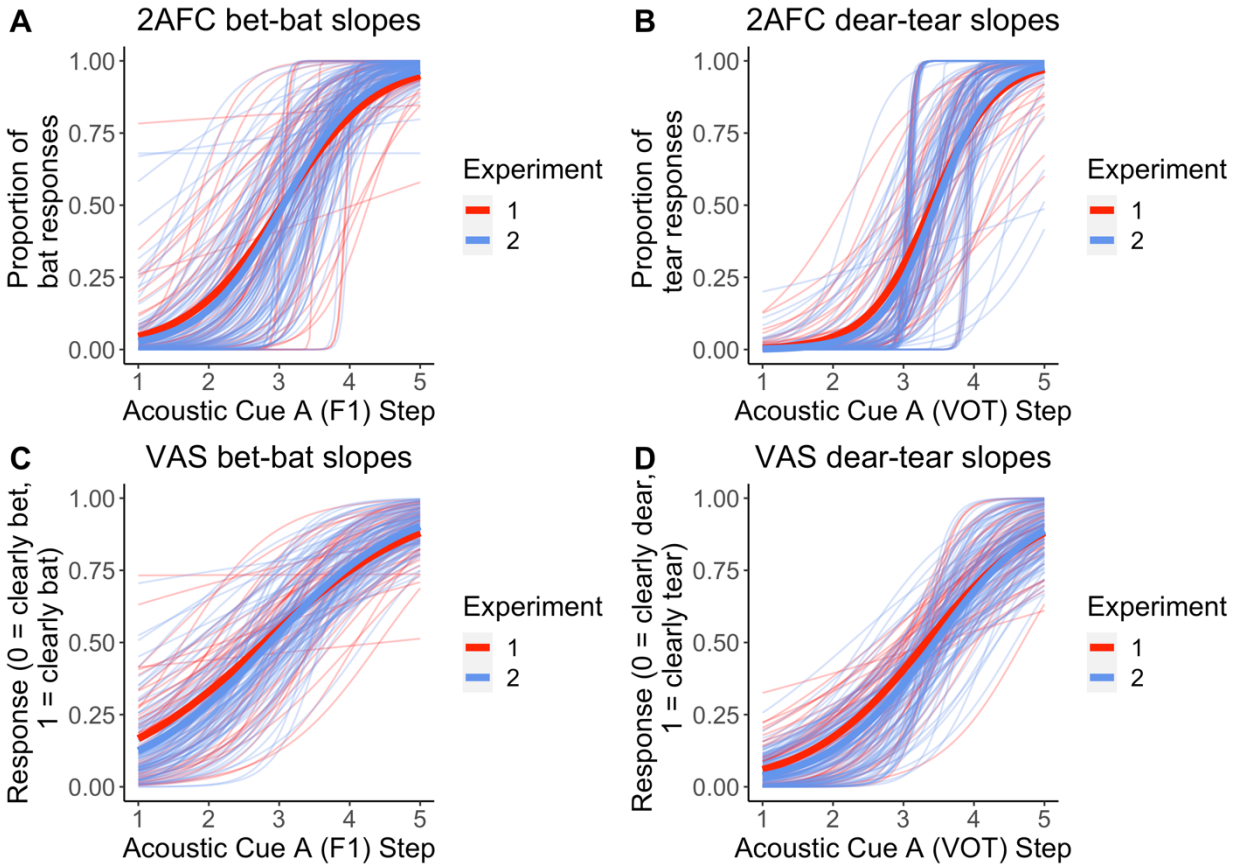
Performance on the two native language perception tasks is shown in Figure 1, and representative individual results are shown in Figure 2. When averaged across all participants, overall response patterns were similar across Experiment 1 and Experiment 2 (compare thick red and blue lines in Figure 1). However, significant individual variability was observed across both tasks and both experiments, with participants showing identification slopes ranging from very shallow to very steep (see thin lines in Figure 1 and example participants in Figure 2). Steeper slopes are more evident on the 2AFC task (no doubt due to its categorical nature) than on the VAS task. Participants also differed in the consistency of their responses, that is, in how closely their response to each stimulus fell around their predicted identification slope (Figure 2).

Violin plots of scores on the non-native perception task and the cognitive tasks are displayed in Figure 3. Overall performance (mean and standard deviation on each task) was very similar for Experiment 1 and Experiment 2, and similar variability in scores was also observed across both experiments as shown by the overlap between red and blue plots. As anticipated, the non-native discrimination task was generally challenging, with mean accuracy falling at 53-54% for both experiments; and at the individual level some participants had particular difficulty discriminating the non-native sounds (accuracy around 25%, at chance), while others were quite successful (accuracy of 75% and above; Figure 3A). Participants also showed a range of scores on the attention and memory tasks (Figure 3B and 3C).

Prior to the main analyses, for the sake of ease of interpretation and exploration of the data, pairwise correlations were computed and visualized between the variables of interest for each hypothesis. These pairwise correlations are provided in Supplemental Figures S.1 to S.4.

Figure 1

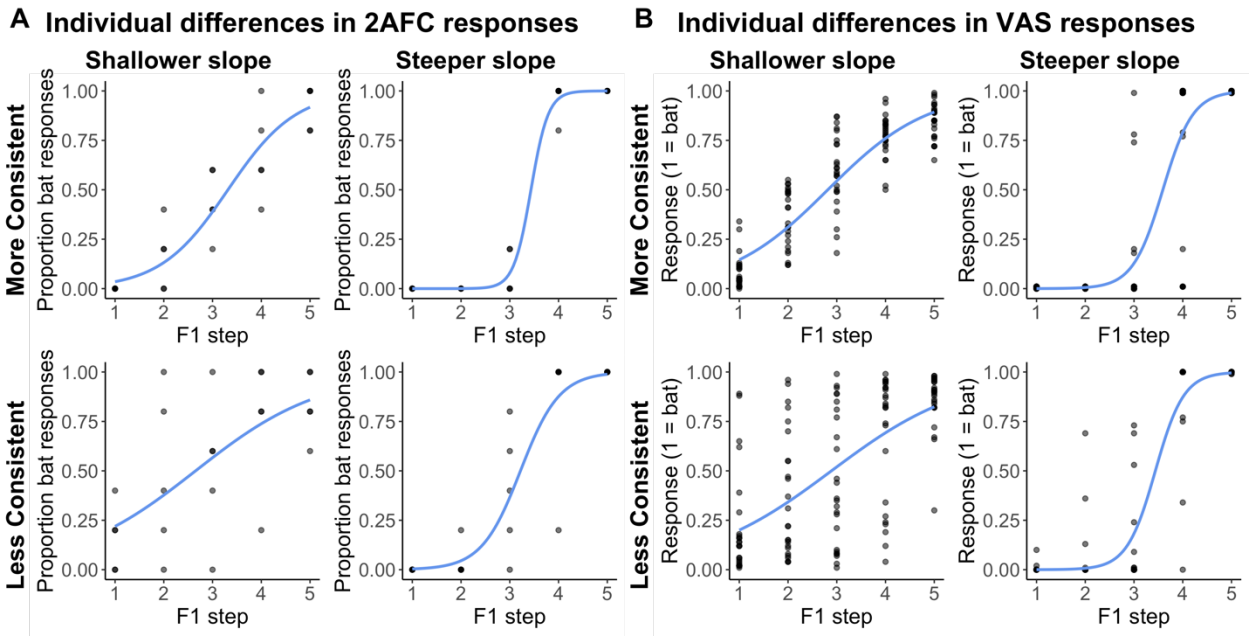
Group and individual responses on the native perception tasks (2AFC and VAS), for both experiments.



Note. (A) 2AFC bet-bat responses by cue A, (B) 2AFC dear-tear responses by cue A, (C) VAS bet-bat responses by cue A, and (D) VAS dear-tear responses by cue A. Thin lines are logistic curves fit to each individual participant for each step of acoustic cue A, and thick lines are logistic curves fit to the whole dataset. VAS responses varied continuously from 0-100, but were transformed to range from 0-1 for the purposes of fitting logistic curves to the data for these plots (regular logistic regression was used here for visualization purposes, rather than the rotated logistic function fit to the VAS data as described in the *Preparatory Data Analysis* section).

Figure 2

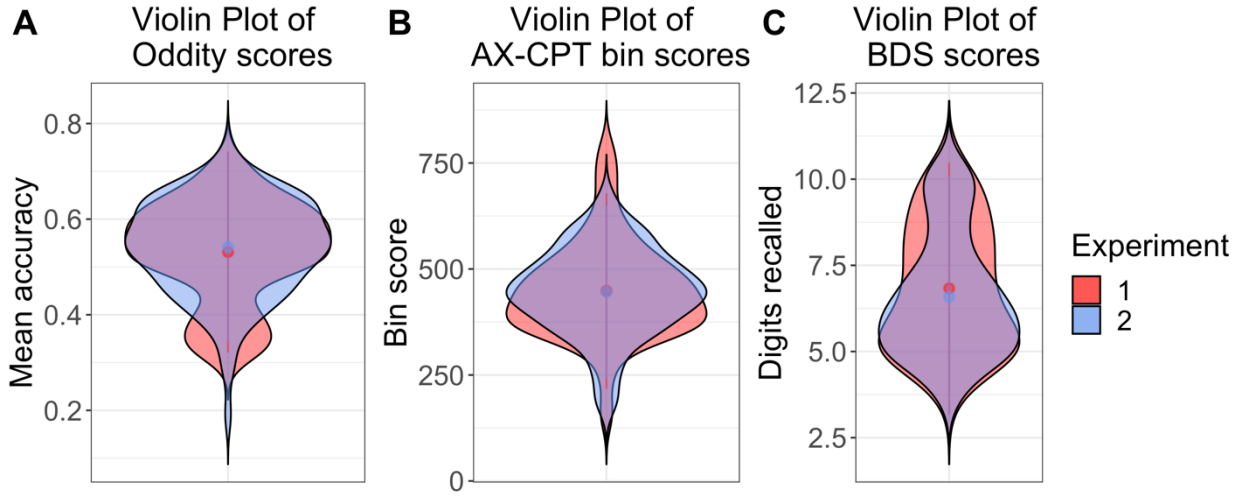
Examples of individual variability on the native perception tasks (2AFC and VAS)



Note. For the 2AFC task (A), each dot is the participant's average response across the five presentations of a given stimulus. For the VAS task (B), each dot is a participant's response on a given trial. Lines are logistic curves fit to responses; dots clustered closely around the fitted curve indicate more consistent responses. VAS responses varied continuously from 0-100, but were transformed to range from 0-1 for the purposes of fitting logistic curves to the data for these plots (regular logistic regression was used here for visualization purposes, rather than the rotated logistic function fit to the VAS data as described in the *Preparatory Data Analysis* section). Top left of each plot: shallow and consistent, top right: steep and consistent, bottom left: shallow and inconsistent, bottom right: steep and inconsistent. The participants in each panel are chosen as representative examples of variability on the task and are not the same across both tasks.

Figure 3

Distributions of performance on non-native and control tasks



Note. (A) Oddity task, (B) AX-CPT task, and (C) Backwards Digit Span task, for both experiments. Mean and standard deviation are indicated by the dot and vertical line within each plot. Purple indicates overlap between the two experiments.

Hypothesis 1 – Canonical Correlation

Analysis

Our first hypothesis was that 2AFC slopes would not relate to VAS slopes but would relate to the consistency of VAS responses, with inconsistent VAS performance predicting shallower 2AFC slopes. Since we had 4 2AFC slope measures, 2 VAS slope measures and 2 VAS consistency measures, we tested our hypothesis by first running canonical correlation analyses, which test the strength of relationships between two sets of variables. Canonical correlation is a dimensionality reduction technique similar to principal component analysis (PCA), but while PCA aims to determine the dimensions that account for the most variance *within* a set of variables, canonical correlation analyses aim to determine the dimensions that account for the most covariance *between* two sets of variables. Canonical correlation analyses output canonical correlation coefficients, which measure the strength of the association between pairs of canonical variates (each pair of canonical variates is called a canonical dimension, so the canonical correlation coefficients can also be thought of as representing the strength of each canonical dimension). A canonical variate is an orthogonal, linear combination of the variables within a set—the variables are weighted so as to maximize the correlation between the canonical variate derived from that set of variables and the canonical variate derived from the other set of variables of interest (i.e., to maximize the correlation coefficient for a given canonical dimension). Canonical variates are latent variables and can be considered analogous to the factors derived from factor analysis. The number of canonical variate pairs or canonical dimensions is equal to the number of variables in the smallest set; in this case, there are two canonical dimensions for each canonical correlation. A significant correlation along one or both

dimensions suggests a relationship between the two sets of variables. Statistical significance of the canonical correlation coefficients for each dimension was evaluated using Wilks' lambda. More information on canonical correlation analysis can be found in Sherry & Henson (2005) and UCLA: Statistical Consulting Group (n.d.).

Canonical correlation 1 was between the four 2AFC random slopes coefficients and the two VAS slope measures. Hypothesis 1 predicted that these sets of variables would not be related. Canonical correlation 2 was between the four 2AFC random slopes coefficients and the two VAS consistency measures. Hypothesis 1 predicted that these sets of variables would be related. These analyses were conducted in R using the packages CCA (Canonical Correlation Analysis; González & Déjean, 2021) and CCP (Significance Tests for Canonical Correlation Analysis; Menzel, 2012). When interpreting effect size of the results, we follow the guidelines established by Gignac & Szodorai (2016) for individual differences research (small: $r = 0.1$; medium: $r = 0.2$; large: $r = 0.3$) and those established by Plonsky & Oswald (2014) for second language research (small: $r = 0.25$; medium: $r = 0.4$; large: $r = 0.6$). As such, a correlation < 0.1 is considered small and > 0.6 is considered large, while intermediate values are referred to by a combination of the two guidelines (e.g., 0.4 is considered medium by Plonsky & Oswald and large by Gignac & Szodorai, so we consider such a value to reflect a medium-large effect size).

Results – Experiment 1

Canonical correlation 1 revealed that the relationship between 2AFC coefficients and VAS slope measures was large and significant along the first canonical dimension ($r_c = 0.690$, $p < 0.001$), and medium-large but did not reach significance along the second canonical dimension ($r_c = 0.439$, $p = 0.037$). The significant correlation along the first dimension suggests that, contrary to our hypothesis, there does appear to be a relationship between the 2AFC coefficients

and VAS slopes. A scatterplot of this significant correlation is displayed in Figure 4A. Note that due to the complex patterns of loadings of the original variables onto the first canonical dimension, this positive canonical correlation coefficient does not indicate that slopes are positively related across the 2AFC and VAS tasks; rather, the relationship between slopes across tasks appears to depend on the contrast and acoustic cue. Furthermore, the result should be treated with caution as the effect is smaller and no longer significant with increased statistical power, as described in the results of Experiment 2 below.

Canonical correlation 2 revealed that the relationship between 2AFC coefficients and VAS consistency measures was large and significant along the first canonical dimension (r_c = 0.617, $p < 0.001$) and medium-large but did not reach significance along the second canonical dimension (r_c = 0.410, p = 0.064). Thus, in line with our hypothesis, there does appear to be a relationship between the 2AFC coefficients and VAS consistency measures. A scatterplot of the significant correlation along the first canonical dimension is displayed in Figure 4B.

For both correlations, Supplemental Table S.2 displays the canonical correlation coefficients and their significance, and Supplemental Table S.3 displays canonical coefficients showing loadings of the original variables onto each canonical dimension. The interpretation of canonical coefficients is analogous to the interpretation of regression coefficients.

Results – Experiment 2

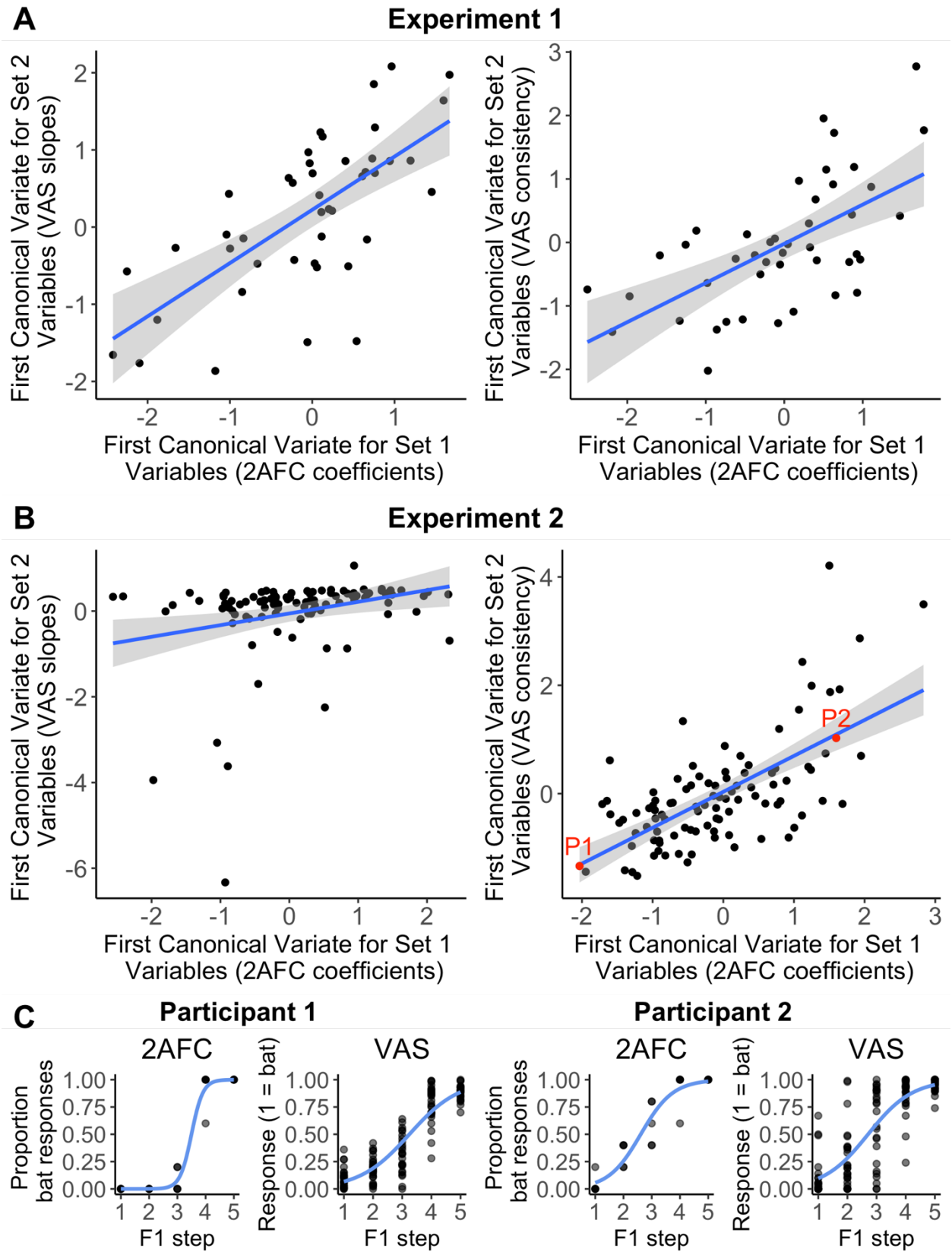
Canonical correlation 1 revealed that the relationship between 2AFC coefficients and VAS slope measures was small-medium and statistically insignificant along the first canonical dimension (r_c = 0.272, p = 0.423) and small and statistically insignificant along the second canonical dimension (r_c = 0.090, p = 0.855). Notice how the effect size is smaller than in Experiment 1. See Supplemental Table S.2 for canonical correlation coefficients and their

significance, and Supplemental Table S.3 for canonical coefficients showing loadings of the original variables onto each canonical dimension. The lack of significant correlation is in line with our hypothesis that 2AFC coefficients and VAS slopes are not related; perhaps the significant correlation found in Experiment 1 was due to an insufficient sample size. A scatterplot of the non-significant correlation along the first canonical dimension is displayed in Figure 4C.

Canonical correlation 2 revealed that the relationship between 2AFC coefficients and VAS consistency measures was large and significant along the first canonical dimension (r_c = 0.663, $p < 0.001$), and medium-large and significant along the second canonical dimension (r_c = 0.398, $p < 0.001$). This effect size is similar to what was found for the same analysis in Experiment 1, and stands in comparison to the small effect size observed for correlation 1 between 2AFC coefficients and VAS slopes. See Supplemental Table S.2 for canonical correlation coefficients and their significance. Thus, across both Experiment 1 and Experiment 2, we find support for our hypothesis that 2AFC coefficients and VAS consistency measures are related. A scatterplot of the significant correlation along the first canonical dimension is displayed in Figure 4D, and two participants from opposite ends of the correlation are highlighted as examples. The response patterns of these two participants on the 2AFC and VAS tasks are shown in Figure 4E. Participant 1 illustrates how people with *steeper* 2AFC slopes tend to have *more consistent* VAS responses. Participant 2 illustrates how people with *shallower* 2AFC slopes tend to have *less consistent* VAS responses. Participant 1 has a steep 2AFC slope and shallow VAS slope whereas Participant 2 has similar slopes across both tasks, demonstrating how slopes on the two tasks do not necessarily relate within participants.

Figure 4

Relationships between variables of interest for hypothesis 1


A
Experiment 1
First Canonical Variate for Set 2 Variables (VAS slopes)
First Canonical Variate for Set 1 Variables (2AFC coefficients)
First Canonical Variate for Set 2 Variables (VAS consistency)
First Canonical Variate for Set 1 Variables (2AFC coefficients)
B
Experiment 2
First Canonical Variate for Set 2 Variables (VAS slopes)
First Canonical Variate for Set 1 Variables (2AFC coefficients)
First Canonical Variate for Set 2 Variables (VAS consistency)
P1
P2
First Canonical Variate for Set 1 Variables (2AFC coefficients)
C
Participant 1
Participant 2
2AFC
VAS
Proportion bat responses
Response (1 = bat)
F1 step

Note. (A) Experiment 1: Scatterplots of the correlation between the first pair of canonical variates, for the canonical correlation between 2AFC coefficients and VAS slopes (left) and the canonical correlation between 2AFC coefficients and VAS consistency (right). *n = 44.* (B) Experiment 2: Scatterplots of the correlation between the first pair of canonical variates, for the canonical correlation between 2AFC coefficients and VAS slopes (left) and the canonical correlation between 2AFC coefficients and VAS consistency (right). Each dot represents a participant, and blue lines are lines of best fit with 95% CIs. Two representative participants from different ends of the correlation are highlighted in red. *n = 100.* (C) Experiment 2: Response patterns on the 2AFC and VAS tasks, for the two representative participants highlighted in (B). For the 2AFC task, each dot is the participant's average response across the five presentations of a given stimulus. For the VAS task, each dot is a response on a given trial. Lines are logistic curves fit to responses; dots clustered closely around the fitted curve indicate more consistent responses. Participant 1 has a steep 2AFC slope, shallow VAS slope, and consistent VAS responses. Participant 2 has shallow 2AFC and VAS slopes, and inconsistent VAS responses. These two participants illustrate how slopes across tasks are not necessarily related within participants, and how steeper 2AFC slopes are associated with more consistent VAS responses. Note that VAS responses varied continuously from 0-100, but were transformed to range from 0-1 for the purposes of fitting logistic curves to the data for these plots (regular logistic regression was used here for visualization purposes, rather than the rotated logistic function fit to the VAS data as described in the *Preparatory Data Analysis* section).

Hypothesis 1 – Multivariate Multiple Regression

Analysis

Following up on the correlations, we conducted a multivariate multiple regression analysis. This enabled us to include all four 2AFC coefficients as the response and all four VAS measures of interest as predictors, as well as attention and memory measures as additional control predictors. In doing so, we were able to determine whether any relationships found through canonical correlation would still hold after controlling for these additional predictors.

Using the lm() function, the model equation in R was: cbind(2AFC bet-bat acoustic cue A slope, 2AFC bet-bat acoustic cue B slope, 2AFC dear-tear acoustic cue A slope, 2AFC dear-tear acoustic cue B slope) ~ VAS bet-bat slope + VAS dear-tear slope + VAS bet-bat consistency + VAS dear-tear consistency + AX-CPT bin score + Digit span level. We then used multivariate tests (Type II MANOVA) to evaluate the significance of each predictor across the four models,

while accounting for the covariances between coefficients. This was done with the Anova() function from the car package in R (Fox & Weisberg, 2019).

Results

Output from the multivariate multiple regression model includes regression tables from four separate regression models, fit with each 2AFC coefficient as the response; this output is shown in Supplemental Tables S.4 and S.5 for Experiment 1 and Experiment 2 respectively. Model validation plots (quantile-quantile plots of residuals, plots of fitted values against residuals, and plots of Cook's distance per participant) can be found in Supplemental Figures S.5 to S.7.

Multivariate tests (Type II MANOVA) were used to evaluate the significance of each predictor across the four models while taking into account the covariances between coefficients (Table 3). These analyses revealed that, in line with our hypothesis and with the canonical correlation results, the VAS consistency measures significantly predicted 2AFC coefficients after accounting for other predictors. The AX-CPT and backwards digit span predictors were not significant. These findings held across both experiments. For Experiment 1, the VAS slope measures significantly predicted the 2AFC coefficients (contrary to our hypothesis); however, with the increased power obtained in Experiment 2, this relationship disappeared.

Table 3

Summary of the multivariate multiple regression model predicting 2AFC coefficients, for each experiment

Experiment 1					
Predictor	Pillai's trace	*F*	Num *df*	Den *df*	*p*
VAS bet-bat slope	0.591	10.854	4	30	<0.001
VAS dear-tear slope	0.305	3.291	4	30	0.024
VAS bet-bat consistency	0.318	3.496	4	30	0.019
VAS dear-tear consistency	0.411	5.239	4	30	0.003
AX-CPT bin score	0.234	2.293	4	30	0.082
Backwards digit span	0.076	0.614	4	30	0.656
n = 40					
Experiment 2					

Predictor	Pillai's trace	*F*	Num *df*	Den *df*	*p*
VAS bet-bat slope	0.049	1.058	4	83	0.383
VAS dear-tear slope	0.370	0.796	4	83	0.531
VAS bet-bat consistency	0.201	5.207	4	83	<0.001
VAS dear-tear consistency	0.206	5.391	4	83	<0.001
AX-CPT bin score	0.039	0.852	4	83	0.497
Backwards digit span	0.078	1.761	4	83	0.145
n = 93					

Note. Model equation: cbind(2AFC bet-bat acoustic cue A slope, 2AFC bet-bat acoustic cue B slope, 2AFC dear-tear acoustic cue A slope, 2AFC dear-tear acoustic cue B slope) ~ VAS bet-bat slope + VAS dear-tear slope + VAS bet-bat consistency + VAS dear-tear consistency + AX-CPT bin score + Backwards digit span.

Hypothesis 2 – Multiple Regression

Analysis

Hypothesis 2 involved predicting non-native perception from all native perception and control measures, which would have resulted in a model with ten predictors. To reduce the number of predictors and thus reduce overfitting while increasing power, dimensionality of the native perception measures was reduced using PCA, as implemented by the prcomp() function in R. The same procedure was followed for both experiments: one PCA was run on the four 2AFC coefficients and another was run on the four VAS variables (two slope and two consistency measures). The first two components from each PCA were then extracted for analysis. Correlations between the original variables and the extracted principal components for both experiments are displayed in Table 4. Across the two experiments, all four 2AFC variables were correlated in the same direction with the first component suggesting that this component reflected 2AFC slopes in general, and bet-bat acoustic cue B was strongly positively correlated with the second component. For the VAS measures across the two experiments, slopes and consistency were correlated in opposite directions with the first component while bet-bat and dear-tear measures were correlated in opposite directions with the second component, suggesting

that the first component distinguishes between slope and consistency while the second one distinguishes between the two contrasts.

In order to test hypothesis 2, a multiple regression model was fit. The response was oddity *A* scores, and the predictors were the first two principal components derived from the PCA of the 2AFC coefficients, the first two principal components derived from the PCA of the VAS measures, and the two control predictors. Because visualization of the distribution of oddity *A* scores for both experiments revealed some negative skew, the scores were exponentially transformed; models were then fit predicting the scores both with and without the transformation, and the model with the best performance is reported. Using the lm() function, the model equation in R was: Oddity A score ~ 2AFC principal component 1 + 2AFC principal component 2 + VAS principal component 1 + VAS principal component 2 + AX-CPT bin score + Digit span level. Hypothesis 2 posited that oddity scores would be predicted by the 2AFC and VAS measures even after accounting for the control predictors.

Results

Hypothesis 2 was not supported; the anticipated predictors did not significantly predict oddity scores. The multiple regression model for Experiment 1 is summarized in Table 5, and model validation plots can be found in Supplemental Figures S.8 to S.10. For Experiment 1, the first principal component derived from the 2AFC measures was a significant predictor; however, with the increased power obtained in Experiment 2, this relationship disappeared. For Experiment 2, none of the predictors was significant (the first principal component derived from the VAS measures showed the largest coefficient but did not reach significance, $\hat{\beta} = 0.042$, $p = 0.076$; see Supplemental Table S.6).

For all of the regression models run for these two experiments, we checked for influential participants as indicated by Cook's distance. In the case of the multiple regression model for Experiment 2, one participant was found to have higher influence than the others (see Supplemental Figure S.11); upon further examination, this participant interpreted the VAS task differently, responding primarily at the endpoints of the slider rather than along its entire range. A model was run excluding this high-influence participant, since this individual did not appear to be representative of the behaviour of our sample. This additional model is summarized in Table 5, along with model validation plots in Supplemental Figures S.8 to S.10. The first principal component derived from the VAS measures—primarily reflecting VAS consistency—was a significant predictor ($\hat{\beta} = 0.096$, $p = 0.002$). The relationship between non-native perception and VAS consistency (averaged across both contrasts) is displayed in Figure 5, revealing how more consistent VAS responses were associated with better non-native discrimination. The original model including the influential participant can be found in Supplemental Table S.6, along with model validation plots in Supplemental Figure S.11.

Table 4

Correlations between the original 2AFC variables and the first two principal components extracted from them (left), and between the original VAS variables and the first two principal components extracted from them (right)

Experiment 1

	PC1	PC2		PC1	PC2
2AFC bet-bat acoustic cue A	0.466	-0.223	VAS bet-bat slope	0.421	-0.076
2AFC bet-bat acoustic cue B	0.290	0.952	VAS dear-tear slope	0.222	0.818
2AFC dear-tear acoustic cue A	0.592	-0.112	VAS bet-bat consistency	-0.593	-0.287
2AFC dear-tear acoustic cue B	0.590	-0.180	VAS dear-tear consistency	-0.650	0.492
Percent variance explained	62%	22%	*Percent variance explained*	36%	30%

Experiment 2

	PC1	PC2		PC1	PC2
2AFC bet-bat acoustic cue A	-0.360	0.537	VAS bet-bat slope	0.399	-0.359
2AFC bet-bat acoustic cue B	-0.169	0.766	VAS dear-tear slope	0.257	0.773
2AFC dear-tear acoustic cue A	-0.647	-0.266	VAS bet-bat consistency	-0.625	-0.321
2AFC dear-tear acoustic cue B	-0.651	-0.232	VAS dear-tear consistency	-0.620	0.414
Percent variance explained	52%	29%	*Percent variance explained*	41%	29%

Table 5

Summary of the multiple regression model predicting oddity A scores, for each experiment

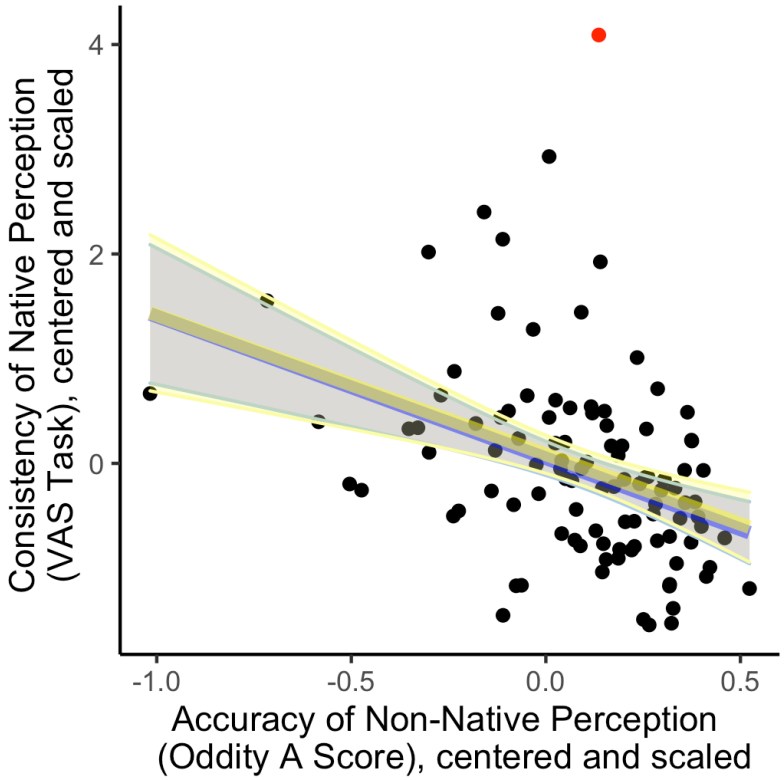
Experiment 1				
Coefficient	$\hat{\beta}$	$SE(\hat{\beta})$	t	p
(Intercept)	1.731	0.047	37.146	<.001
2AFC principal comp. 1	0.071	0.033	2.138	0.040
2AFC principal comp. 2	-0.038	0.050	-0.754	0.456
VAS principal comp. 1	0.033	0.042	0.782	0.440
VAS principal comp. 2	-0.034	0.044	-0.779	0.441
AX-CPT bin score	-0.022	0.052	-0.436	0.666
Backwards digit span	0.030	0.050	0.607	0.548
Multiple R^2 = 0.233, Adjusted R^2 = 0.094, Residual SE = 0.292 (df = 33), n = 40				

Experiment 2				
Coefficient	$\hat{\beta}$	$SE(\hat{\beta})$	t	p
(Intercept)	0.079	0.025	3.153	0.002
2AFC principal comp. 1	-0.010	0.021	-0.452	0.653
2AFC principal comp. 2	-0.002	0.023	-0.073	0.942
VAS principal comp. 1	0.096	0.030	3.257	0.002
VAS principal comp. 2	-0.021	0.028	-0.737	0.463
AX-CPT bin score	-0.053	0.028	-1.896	0.061
Backwards digit span	-0.056	0.029	-1.925	0.058
Multiple R^2 = 0.211, Adjusted R^2 = 0.155, Residual SE = 0.232 (df = 84), n = 91				

Note. Model equation: Oddity A score (exponentially transformed for Experiment 1 but not for experiment 2, based on comparisons of model performance) ~ 2AFC principal component 1 + 2AFC principal component 2 + VAS principal component 1 + VAS principal component 2 + AX-CPT bin score + Backwards digit span.

Figure 5

Relationship between non-native discrimination and native VAS consistency



Note. Data are from Experiment 2. Higher values indicate less consistency and better non-native perception. Each dot represents a participant, with the outlier excluded from analyses in red. In blue is the line of best fit with 95% CI when the outlier is excluded, and in yellow is the line of best fit with 95% CI when the outlier is included.

Non-Preregistered Analyses

In addition to the analyses that were preregistered on the OSF, a variety of additional analyses were run. The details of all of these analyses can be found in the R Markdown document on the OSF. Together with the preregistered analyses, these analyses provided a more in-depth understanding of how individual variability is structured across the two tasks.

Comparing Slope Estimate Methods

In our analyses above, as in Kapnoula et al. (2017), we found that the 2AFC slopes and the VAS slopes were not correlated across individuals, which seems to indicate that they are not measuring the same aspect of performance. However, as discussed in the introduction, they are using different methods to measure slope, and thus they might not be directly comparable. We

therefore extended the work of Kapnoula et al. (2017) by fitting their rotated logistic function to the 2AFC data (as was done in Ou et al., 2021) as well as to the VAS data. This enabled a more direct comparison of slopes across the two tasks. Because we had more than one slope variable per task, we compared slope estimates across tasks using canonical correlation. As mentioned in the *Analysis* section for hypothesis 1, canonical correlation evaluates the strength of relationships between two sets of variables and outputs canonical correlation coefficients representing the strength of canonical dimensions. Canonical dimensions are combinations of the original sets of variables, weighted in such a way as to maximize the correlation between sets. Canonical correlation revealed that 2AFC slopes were significantly related across the two calculation methods (mixed-effects logistic regression vs. rotated logistic function), with a large effect size for Experiment 1 ($r_c = 0.63$, $p < 0.001$ for the first canonical dimension) and medium-large effect size for Experiment 2 ($r_c = 0.38$, $p < 0.005$ for the first canonical dimension). Further canonical correlations were then used to determine whether the new 2AFC rotated logistic slopes related to VAS slopes and consistency in similar ways to the original 2AFC mixed-effects regression slopes. These analyses revealed that the relationship between 2AFC rotated logistic slopes and VAS slopes was small to small-medium and did not reach significance, in line with the results from our preregistered analyses (first canonical dimension: $r_c = 0.23$, $p = 0.92$ for Experiment 1, and $r_c = 0.08$, $p = 0.59$ for Experiment 2). This means that the different ways of measuring slope in the two tasks cannot account for the lack of evidence for a relationship between them. We note that, unlike the mixed-effect regression slopes, the 2AFC rotated logistic slopes showed a small and statistically insignificant relationship to VAS consistency (first canonical dimension: $r_c = 0.05$, $p = 0.88$ for Experiment 1, and $r_c = 0.18$, $p = 0.33$ for Experiment 2). However, they do pattern together in the PCA analysis discussed in the *Dimensionality Reduction* section below.

Relating Slopes and Consistency Within 2AFC and VAS Tasks

In our preregistered analyses we compared the predictability of slopes versus consistency measures and in Figure 2 we illustrated examples of participants with all four combinations of high and low consistency and steep and shallow slopes. However, we don't know to what extent these measures are independent of each other. It could be the case that gradient perception facilitates highly consistent responses through providing a detailed and accurate phonetic representation. On the other hand, it could be that those who tend to use just the endpoints of the continuum are the most consistent. These two possibilities would give very different interpretations to the consistency measure. To better understand response consistency, canonical correlations were used to examine the relationship between slopes and consistency *within* each task. These correlations revealed that the relationship between 2AFC slopes (as calculated by the rotated logistic) and 2AFC consistency was large and significant for Experiment 1 (first canonical dimension: $r_c = 0.68$, $p < 0.001$) and medium-large and significant for Experiment 2 (first canonical dimension: $r_c = 0.38$, $p < 0.001$). Similarly, the relationship between VAS slopes and VAS consistency was medium-large and significant for Experiment 1 (first canonical dimension: $r_c = 0.44$, $p = 0.017$) and large and significant for Experiment 2 (first canonical dimension: $r_c = 0.57$, $p < 0.001$). Specifically, steeper 2AFC slopes were associated with more consistent 2AFC responses, and shallower VAS slopes were associated with more consistent VAS responses. This is an important observation that we will return to in the discussion.

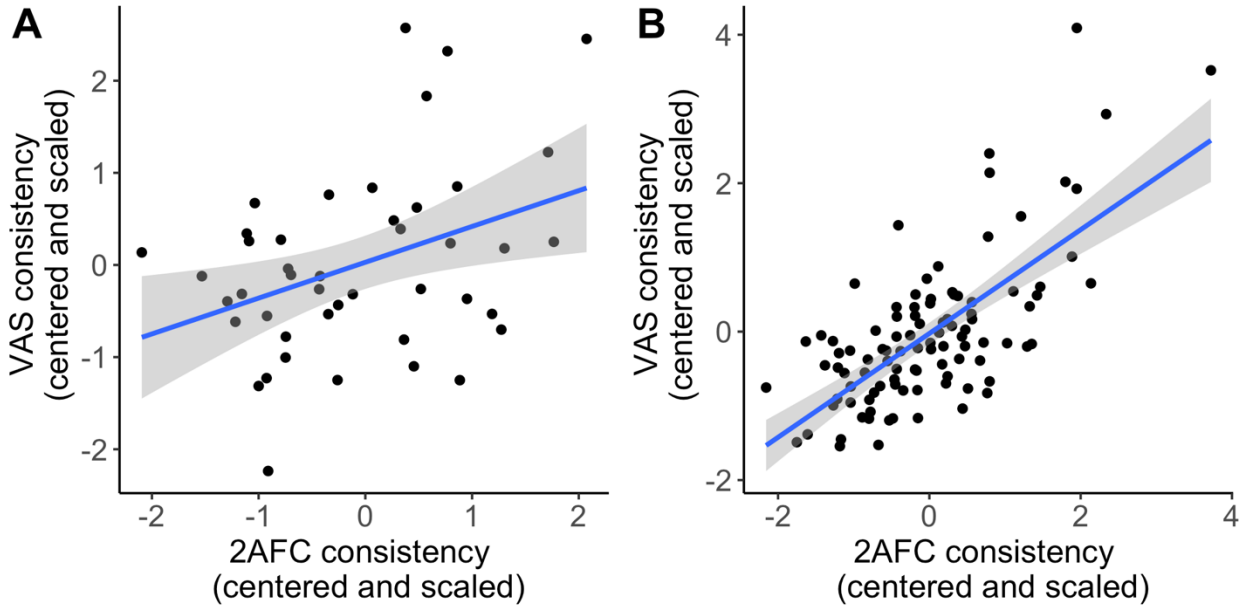
Relating Consistency Across Tasks

Since we now had consistency measures for both tasks, we also examined the relationship between consistency across tasks. Using canonical correlation, we related 2AFC consistency to VAS consistency in order to determine whether some individuals generally show more consistent

phonetic perception than others. This analysis showed that the relationship between consistency on the two tasks was large and significant along the first canonical dimension ($r_c = 0.58$, $p < 0.001$ for Experiment 1, and $r_c = 0.70$, $p < 0.001$ for Experiment 2) and medium-large and significant along the second canonical dimension ($r_c = 0.34$, $p = 0.018$ for Experiment 1, and $r_c = 0.41$, $p < 0.001$ for Experiment 2), suggesting a robust relationship. Figure 6 displays the significant relationship between 2AFC and VAS consistency for both experiments.

Figure 6

Relationship between 2AFC consistency and VAS consistency



Note. Data are presented from Experiment 1 (A) and Experiment 2 (B). Consistency is averaged across the two contrasts presented in the experiments, and higher values indicate less consistency. Each dot represents a participant, and the blue line is a line of best fit with 95% CI.

Dimensionality Reduction of 2AFC and VAS Variables

The above analyses suggest that shallow VAS slopes, steep 2AFC slopes (measured by mixed-effect logistic regression), and consistent responses all pattern together across individuals. Our final analysis confirmed this overall picture by putting all 12 variables (two VAS slopes, two VAS consistency measures, four 2AFC mixed-effects regression slopes, two 2AFC rotated logistic slopes, and two 2AFC consistency measures) into a PCA analysis to see how well they

could be reduced to a smaller set of dimensions. Correlations between the original variables and the first five principal components derived from them are displayed in Supplemental Table S.7, and biplots are displayed in Supplemental Figure S.12. We found that the first principal component was made up primarily of the four consistency measures and the slope measures from the mixed-effect logistic regression of the 2AFC task (with opposite signs from the consistency measures). This confirms that differences in consistency (Figure 6) and their relationship to categorization steepness (right side of Figure 4) capture the greatest amount of variability between individuals. The second principal component shows a similar pattern, with the mixed-effect slopes for the 2AFC task patterning opposite to all of the rotated logistic values (including slope this time as well as consistency for both tasks). The second component also reflects a distinction between the two contrasts, as the bet-bat and dear-tear measures have different signs. Thus, the PCA analysis confirms the patterns observed in the previous canonical correlation analyses and provides a coherent picture of the structure of individual variability in these tasks.

Discussion

The objectives of the current studies were twofold. First, we aimed to clarify whether responses on 2AFC and VAS tasks reflect distinct individual differences in native speech sound perception, with VAS slopes relating to gradiency and 2AFC slopes relating to consistency. We compared participants' responses to identical continua of stimuli on a 2AFC and a VAS task and found that there was no evidence for a relationship between 2AFC identification slopes and VAS identification slopes, but there was a relationship between 2AFC identification slopes and the consistency of VAS responses. Thus, for the first time the findings clearly show that the two tasks measure separate constructs: 2AFC slopes tap into the consistency of perception, while VAS slopes tap into the gradiency of perception.

Second, we aimed to determine whether discrimination of difficult non-native contrasts could be predicted by differences in native phonetic perception as measured by 2AFC and VAS tasks. While we did not find evidence for a relationship between gradiency and non-native perception, we found preliminary evidence that consistent native perception may play a role in discriminating unfamiliar language sounds.

Identification Slopes on 2AFC and VAS Tasks Reflect Different Constructs

Recall that there is ambiguity as to what 2AFC slopes represent, since it is unclear whether a participant with a shallow 2AFC slope (1) has underlyingly gradient perception and is responding probabilistically across trials, or (2) is responding inconsistently across trials. VAS tasks can disambiguate the constructs of gradiency and consistency, and so by comparing VAS performance to 2AFC performance we can determine how the tasks are related and which individual differences each one seems to be measuring.

Based on a marginal relationship between 2AFC slopes and consistency of VAS responses, Kapnoula et al. (2017) proposed that 2AFC and VAS tasks assess different aspects of speech perception. We hypothesized that the two tasks do indeed measure distinct constructs—with 2AFC slopes largely reflecting consistency of perception and VAS slopes largely reflecting gradiency of perception—and that this result might emerge more clearly with some methodological modifications and a large sample size. Instead of presenting continua with different numbers of steps on the two tasks as in Kapnoula et al. (2017), we used exactly the same stimuli in both tasks to facilitate comparison of performance across tasks. Our stimuli included both vowels and consonants rather than consonants alone, increasing the generalizability of the results. We also derived 2AFC and VAS slopes both using different analysis methods (by-participant random slopes from mixed-effects logistic regression vs. slopes

from a rotated logistic function developed by Kapnoula et al. (2017)) and using the same rotated logistic function across tasks, which enabled a more direct comparison of slopes than in previous studies.

It was important to conceptually replicate Kapnoula et al. (2017)'s work in order to advance the field by determining which individual differences are measured by different tasks. The relationship that they reported between 2AFC slopes and VAS consistency could have been spurious, especially given that it was marginal; if the two measures were in fact not related, this would leave us without an understanding of what 2AFC slopes are truly measuring (not consistency or gradiency, but some other construct). On the other hand, if a clearer relationship did emerge between 2AFC slopes and VAS consistency after the implementation of a few methodological changes, this would imply differences in what each task is measuring and would have repercussions for speech perception researchers in terms of which tasks and measures to employ.

In Experiment 1, we found that 2AFC slopes related to VAS consistency as hypothesized, but unexpectedly they also related to VAS slopes. This finding may have been a spurious one due to limited power, because when re-running the analyses with a larger sample size in Experiment 2, we found evidence for a relationship between 2AFC slopes and VAS consistency but not between 2AFC slopes and VAS slopes, in line with our hypothesis. Importantly, in both studies, the relationships between 2AFC slopes and VAS consistency were statistically significant (not only marginal as had previously been found), showing replicability of this finding. These relationships also held across both studies after taking into account individual differences in attention and working memory.

It could be argued that the lack of evidence for a relationship between 2AFC and VAS slopes in Experiment 2 was due in part to the different methods used to calculate slopes on each task (regular logistic mixed-effects regression for the 2AFC task vs. rotated logistic function for the VAS task). In order to provide a more direct comparison of slopes across tasks, our non-preregistered analyses involved fitting the rotated logistic function from Kapnoula et al. (2017) to both 2AFC and VAS data. This approach enabled us to derive slope and consistency measures in the same way for both tasks, yielding insight into the relationships between slopes and consistency across tasks. Even when calculated using the same rotated logistic method, there was no evidence for a relationship between slopes on the 2AFC and VAS tasks. This finding is also striking given that participants were responding to identical stimuli in both tasks. These analyses provide further evidence that 2AFC and VAS slopes reflect different constructs, strengthening the findings from our preregistered analyses.

The fact that all participants completed the VAS task prior to the 2AFC task (following the procedure described by Kapnoula et al., 2017) could potentially be viewed as a limitation due to the possibility that participants adapted to the stimuli from one task to the next. In their work which measured lexical effects on speech perception over two sessions, Giovannone & Theodore (2023) found that participants showed a weakened Ganong effect from the first to the second session, suggesting increased reliance on acoustic-phonetic information and decreased reliance on lexical information over time (though this was not the case for other tasks such as phoneme restoration). If listeners do indeed tend to increase their reliance on acoustic-phonetic information the more they are exposed to stimuli, this could potentially affect performance on our native perception tasks. However, this possibility would be more of a concern if the 2AFC task had been completed before the VAS task; in that case, increased acoustic-phonetic reliance

during the second task could have resulted in more gradient response tendencies and therefore shallower VAS slopes (although this effect might have been counteracted by the categorical 2AFC task which could bias participants to mainly respond at the VAS slider's endpoints instead of along its whole length). In the case of the 2AFC task being completed second, greater gradiency/reliance on acoustic-phonetic information should not affect responses because we have found 2AFC slopes to be reflective of response consistency rather than of gradiency. Thus, we maintain that the choice to always present the VAS task before the 2AFC task was a theoretically and methodologically sound one.

Individual Differences in the Consistency of Native Perception

The present work clarifies that shallow 2AFC slopes appear to reflect inconsistent rather than gradient perception. This finding is in line with recent electrophysiological work that related participants' 2AFC slopes to measures of their subcortical and cortical auditory encoding (Ou & Yu, 2022). The researchers found that participants with less faithful subcortical encoding of speech had shallower 2AFC slopes, supporting the notion that 2AFC slopes reflect inconsistency in perception (Ou & Yu, 2022). Additionally, our results shed light on why previous studies have suggested an association between shallow 2AFC slopes and language impairment—such an impairment appears to be accompanied by inconsistency or imprecision in perception. This conclusion is further supported by work showing that children with developmental dyslexia, who are known to have shallower 2AFC slopes (Manis et al., 1997; Joanisse, et al., 2000; Serniclaes et al., 2001), also have inconsistent or atypical neural representation of speech (Destoky et al., 2020; Keshavarzi et al., 2022; Power et al., 2016).

Interestingly, our non-preregistered analyses revealed that participants' response consistency values (as extracted from the rotated logistic function from Kapnoula et al. 2017)

were related across the 2AFC and VAS tasks. Response consistency also patterned together across the vowel and consonant contrasts in the PCA analysis. These findings suggest that consistency may be a stable and task-independent property of the individual. While previous work has demonstrated individual differences in consistency on VAS tasks (Kapnoula et al., 2017), we provide evidence that these differences seem to hold across tasks. That is, some listeners appear to be more consistent than others in how they map perceived speech sounds to response options, regardless of the specific format of the response options. An interesting topic for future study could be how and why consistency may reflect optimal perception, as well as the extent to which differences in consistency of perception generalize to other tasks (e.g., speech-in-noise perception, assimilation of non-native sounds to native categories) and other modalities (e.g., ratings of colour stimuli).

An important question for future research is at which level the consistency measured by phonetic perception tasks arises (i.e., whether it is somewhere along the perceptual pathway and/or during higher-level decision-making processes). The work by Ou & Yu (2022) suggests that early subcortical auditory encoding of sound is a source of consistency, but they also found that steep slopes on a 2AFC task were further related to a difference in the representation between cortical and subcortical encoding. This seems to indicate that steep slopes require accurate gradient encoding and consistent transformation into categories. This suggests that perhaps consistency at higher levels of perception and cognition may play an additional role in predicting individual differences in responses, for example through attention or memory. It would also be of interest to investigate whether atypical and typically-developing populations show similar or different sources of inconsistency.

Consistency and Gradiency as Distinct Yet Related Constructs

Separately from consistency, gradiency of perception seems to be its own construct that is best measured by VAS tasks and that may be adaptive (rather than suggestive of an impairment) in various situations as described further below. As discussed by Kapnoula et al. (2017), gradiency and consistency may be orthogonal, and VAS tasks are useful precisely because they enable researchers to calculate a separate measure of each construct. A conceptual distinction between consistency and gradiency makes sense given recent evidence that measures of the two constructs (as extracted from a VAS task) have separate structural correlates in the brain (Fuhrmeister & Myers, 2021). Thus, there appears to be a difference between distinguishing gradual changes along a continuum (gradiency) and having highly reliable mapping between stimulus and response (consistency). This being said, our results do suggest that the constructs relate to one another. We found that listeners with more consistent responses tended to have steeper slopes on the 2AFC task and shallower slopes on the VAS task. This outcome probably reflects an optimal pattern of perception whereby the listener shows categorical responses when presented with a categorical task and gradient options when presented with a gradient task. The most successful listeners therefore appear to be the ones who are consistently able to map their percept to a response option, which promotes precise and optimal responding across tasks.

Although the mechanisms of both consistency and gradiency remain to be elucidated, based on existing work we can speculate that they may have partially distinct and partially overlapping underpinnings which could explain our findings. Behavioural response consistency may arise at least to some extent from neural response consistency, which can be quantified as the similarity of the evoked neural response across repeated presentations of a sound (Krizman & Kraus, 2019; Ou & Yu, 2022). Differences in gradiency may also arise partly from this same neural response consistency, with more similar neural responses promoting more gradient

perception by facilitating the faithful encoding of subtle differences between stimuli; but gradiency may additionally result from the transformations that the neural response undergoes as it travels up the auditory pathway from the brainstem to higher-level cortex, with greater transformation leading to less gradient perception (Ou & Yu, 2022)—this process is referred to as perceptual warping by Kapnoula et al. (2021). Under this possibility, gradiency and consistency share some basic mechanisms and would relate to each other as found in the present work; yet two listeners with equally consistent neural and behavioural responses could still differ in gradiency based on how their neural responses were transformed along the auditory pathway. Nevertheless, this explanation remains purely theoretical, and future work with neural measures will be needed to determine the precise origins of both constructs and to untangle the nature of the relationship between them.

Beyond Categorical Perception and Categorical Tasks

The current findings add to the growing conviction that psycholinguistics should move beyond a purely categorical view of phonetic perception (e.g., Holt & Lotto, 2010; Kapnoula et al., 2017; McMurray, 2022; McMurray et al., 2002; Schouten et al., 2003). We support the view that gradiency can be a beneficial (not suboptimal) strategy during perception (Clayards et al., 2008; Desmeules-Trudel & Zamuner, 2019; Kapnoula et al., 2021). In fact, both categorical and gradient modes of perception are likely to be useful in their own way: the ability to fit sounds into one category or another appears to be an important part of processing sounds efficiently (e.g., Shen & Froud, 2016), and the ability to distinguish within-category differences seems to promote flexibility during perception (e.g., Kapnoula et al., 2021). A given listener's sensitivity to between- versus within-category differences in speech sounds likely depends on idiosyncrasies of their perceptual systems (Kapnoula & McMurray, 2021) and varies according

to the particular context (e.g., more between-category sensitivity when listening to predictable native input that easily fits one's preestablished categories, and more within-category sensitivity when listening to accented or non-native speech that requires perceptual flexibility). In other words, listeners may use different strategies—of which gradiency is one—to arrive at the common goal of deriving concrete representations from the continuous speech signal. Beyond its role in encouraging flexible phonetic perception, gradiency is also no doubt important for the perception of various social factors related to a given speaker, such as emotion (Cowen et al., 2019), geographic dialect (Plichta & Preston, 2005), and perceived masculinity/femininity (Munson, 2007). Considering that all of these social factors exist along continua, it is logical that perceiving them in an accurate and nuanced way would require gradient acoustic representations. The potential sources and functions of gradiency, as well as the context-dependent ways in which it may be combined with other strategies during the perception of speech and of speakers, are pertinent questions to continue exploring with future research.

Our findings have important methodological implications in psycholinguistics and related fields. Notably, researchers should select tasks carefully based on the constructs that they wish to measure, while taking into account the limitations and demands of different tasks. When looking to study gradiency, VAS tasks (with their continuous gradient of response options) are a much more appropriate choice than 2AFC tasks (see Apfelbaum et al., 2022 for further discussion of VAS tasks and their utility). The development and adoption of tasks that encourage more fine-tuned and gradated responses, such as VAS and magnitude estimation tasks (Sprouse, 2007), seem to be an important step in advancing psycholinguistic research by revealing nuances of human perception and cognition that may not otherwise be captured by tasks with limited response options.

Predictors of Non-Native Perception

We hypothesized that shallow VAS slopes and steep 2AFC slopes might both be indicative of the ability to make accurate and fine-tuned judgments about acoustic cues and might therefore relate to better non-native discrimination. This hypothesis was not supported. With our preregistered analyses, we found that performance on the non-native perception task was not robustly predicted by any of the native perception measures across our two studies. However, additional analysis excluding an influential participant in Experiment 2 revealed a potential relationship between VAS consistency and non-native perception. This relationship held even after accounting for non-linguistic cognitive factors (attention and working memory). This finding implies that in order to successfully distinguish new speech sounds, an important underlying factor is not so much the exact nature of native speech sound representations (categorical/gradient), but rather the similarity of these representations across time. While not anticipated, such a link between consistent native perception and accurate non-native perception is reasonable when considered in the context of the latest literature on non-native perception.

Very recent work by Fuhrmeister et al. (2023) is in line with our findings. Similarly to us, the authors hypothesized that more gradient VAS slopes on a native phonetic perception task would relate to better discrimination of non-native phonemes; and yet they found that more *consistent* VAS responses related to better non-native discrimination. In addition, preliminary work by Kapnoula & Samuel (2023) has revealed the same pattern of results: better non-native perception was predicted by more consistent VAS responses rather than by more gradient VAS slopes. Across our experiments and other recent research, the same picture is therefore emerging: in order to discriminate non-native speech sounds, it appears to be helpful to have a strong link between a stimulus and one's response to it. As mentioned above, the level at which such

consistency emerges remains to be clarified. Consistency in auditory brainstem responses relates to preliteracy skills (Bonacina et al., 2021; White-Schwoch et al., 2015) and to phonetic discrimination (Tecoulesco et al., 2020), so it is possible that consistency begins playing a role at the level of early neural encoding and is an important element of native and non-native language acquisition.

Further insight comes from studies that have asked participants to listen to native sounds and assimilate them to non-native categories. Such studies have shown that the ability to consistently map a given non-native phoneme to a particular native category is related to having greater non-native perceptual proficiency (i.e., patterns of acoustic cue weighting during non-native perception that more closely resemble those of native speakers; Kang & Schertz, 2021), a larger non-native vocabulary (Bundgaard-Nielsen et al., 2011), and more extensive experience with the non-native language (Levy, 2009). It therefore seems that consistency of phonetic perception can predict various outcomes of non-native language learning success.

While we are unaware of any existing theories which might explain the precise nature of the relationship between native perceptual consistency and non-native perceptual success, the category precision hypothesis of the revised Speech Learning Model (SLM-r; Flege & Bohn, 2021) addresses a similar relationship in the context of speech production rather than perception. According to this hypothesis, the more consistent and precise a person's native categories are (in this case, consistency being defined as low acoustic variability across multiple productions of a phoneme), the better the person will be at distinguishing new non-native sounds and establishing categories for them. Based on our findings, it is conceivable that a similar hypothesis might apply in the realm of perception, where listeners with more consistent and precise native perception can more readily perceive differences between non-native sounds.

Although our preregistered analyses revealed a somewhat surprising lack of evidence for a relationship between native and non-native perception, similar findings have been reported in the past. For instance, other work has found that gradiency of native perception on a VAS task did not relate to non-native discrimination ability (Fuhrmeister et al., 2023) or to scores on a standardized non-native proficiency task (Kong & Kang, 2022). It may be that native gradiency does not relate strongly to non-native outcomes due to differences in some of the processing strategies involved. This possibility is supported by work showing that native and non-native listeners rely on different strategies—namely, lexical knowledge vs. acoustic cues—during word segmentation (Mattys et al., 2010). It has also been found that native speakers show gradient integration of phonetic information (as measured by eye-tracking) during word recognition, whereas non-native speakers show a categorical pattern (Desmeules-Trudel, 2018). An additional possibility is that greater sensitivity to native speech sounds does promote better non-native perception, but that this relationship emerges later in life. In line with this, Kalaivanan et al. (2023) recently found that for older adults, native perceptual sensitivity (as measured by a gating task) was a robust predictor of non-native discrimination; but for younger adults, general intelligence was a stronger predictor. Perhaps younger adults (like the participants in the present experiments) rely more on fluid cognitive factors including attention and memory, while older adults rely more on crystallized factors including their knowledge of native phonemes (Spreng & Turner, 2019). Future work with older populations could clarify this possibility.

The lack of evidence for a strong relationship observed between our native and non-native perception measures could also be due in part to differences in the tasks used to derive the measures. As an example, on the native perception tasks the stimuli had been manipulated to form a continuum, and each trial involved the presentation of one stimulus; on the non-native

perception task the stimuli were not manipulated, and each trial involved the presentation of three stimuli. The fact that we do see some relationships between native and non-native performance despite the differences in tasks suggests that the relationship may be even more robust when more similar measures are used. Future work could compare native and non-native performance more directly, for example by training non-native perception in advance so that participants can respond to non-native sounds on 2AFC and VAS tasks, or by measuring both native and non-native perception using oddity tasks.

It is also worth pointing out that some of the individual variability observed on our native and non-native perception tasks could have arisen from differences in participants' sociolinguistic knowledge and/or labelling strategies. For instance, the *bet-bat* contrast that we tested here is known to participate in ongoing sound change processes such as the Northern Cities Vowel Shift (McCarthy, 2011) and the California Vowel Shift (D'Onofrio et al., 2019). Given that our participants were recruited from across North America, their varied sociolinguistic knowledge could have contributed to some of the differences in performance observed on the native perception tasks. In the future, it would be interesting to measure sociolinguistic factors and relate them to the kinds of individual differences observed here. Additionally, performance on the non-native perception task could have been influenced by whether participants treated the speech sounds as entirely unfamiliar or as better/worse exemplars of native sounds. As an example, a participant that perceived German /ç/ as a new and unfamiliar sound may have been more successful on our task compared to one that perceived /ç/ as a bad exemplar of English /ʃ/ and consequently assimilated /ç/ and /ʃ/ to the same category. In accordance with this possibility, Mayr & Escudero (2010) have shown that participants who assimilated German contrasts to a single English category (rather than to two distinct categories)

had more difficulty identifying those contrasts. By incorporating other tasks where non-native sounds must be labelled or rated for category goodness, future studies could uncover more nuances of the factors relating to individual differences in non-native perception. Note that we do not view these possible sources of variability in native and non-native perception as limitations; while they may have contributed to the variation in performance that we observed, they do not invalidate the relationships we found.

The present work and work by Fuhrmeister et al. (2023) suggest that non-native perception is predicted by the consistency of native perception. If this finding continues to be replicated, it could provide an exciting avenue for further exploration. For instance, perhaps native perception tasks could be administered as brief pre-screenings in language learning settings as a means of identifying people who would benefit from greater support during the learning process. In any case, an important topic that remains to be addressed is why healthy young adults show such variability in their ability to discriminate non-native phonemes. Work in this area is particularly relevant given that successful phonetic perception appears to be a precursor for language learning more generally, predicting outcomes such as non-native vocabulary learning and reading comprehension (Jakoby et al., 2011; Silbert et al., 2015).

Conclusion

In summary, we demonstrated that identification slopes on 2AFC and VAS tasks do not measure the same individual differences in phonetic perception. While shallow VAS slopes seem to reflect gradient perception that involves fine-tuned sensitivity to within-category differences, shallow 2AFC slopes seem to reflect inconsistent perception. This is important given that 2AFC tasks have been extensively employed in previous work and that their slopes have been thought to support the theory of categorical perception, when in fact the slopes were not necessarily

measuring what researchers intended. This work points to the necessity of accounting for task demands during research and of revising theoretical views in light of new evidence. We join other researchers in recommending the use of VAS tasks rather than commonly used 2AFC tasks in psycholinguistic research, and in encouraging views of phonetic perception that account for within-category sensitivity (Apfelbaum et al., 2022; Kapnoula et al., 2017; McMurray et al., 2002; Munson et al., 2017).

Our analyses also pointed to the construct of consistency as a fruitful subject for future investigation. We found that consistency of responses was related across the 2AFC and VAS tasks, suggesting that it may be a stable property of the individual. We further found that consistent responses were associated with steeper 2AFC slopes and shallower VAS slopes. This pattern seems to indicate that people who can consistently associate a given stimulus with a response show the most optimal pattern of perception across tasks (categorical responses on the 2AFC task, gradient responses on the VAS task).

Finally, we found preliminary evidence that successful non-native phonetic perception may be predicted by the consistency of VAS responses. In the future, this could lead to the development of personalized methods of assisting adult language learners based on their individual perceptual and cognitive profiles. The potential benefits of personalized approaches to learning become evident when considering the notable individual differences in performance that are observed across various phonetic perception and cognitive tasks, both here and in other work (e.g., Golestani & Zatorre, 2009; Hanulíková et al., 2012; Hattori & Iverson, 2010; Lee, 2016; Linck & Weiss, 2015). Furthermore, optimizing language learning in adults is particularly relevant in today's highly diverse and interconnected world, in which learning new languages has become key for many people's social integration and advancement.

References

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

Apfelbaum, K. S., Kutlu, E., McMurray, B., & Kapnoula, E. C. (2022). Don't force it! Gradient speech categorization calls for continuous categorization tasks. *The Journal of the Acoustical Society of America*, *152*(6), 3728-3745. https://doi.org/10.1121/10.0015201

Barch, D. M., Berman, M. G., Engle, R., Jones, J. H., Jonides, J., MacDonald III, A., ... & Sponheim, S. R. (2009). CNTRICS final task selection: working memory. *Schizophrenia bulletin*, *35*(1), 136-152.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software,67*(1). https://doi.org/10.18637/jss.v067.i01.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13-34). Amsterdam: John Benjamins.

Bonacina, S., Huang, S., White-Schwoch, T., Krizman, J., Nicol, T., & Kraus, N. (2021). Rhythm, reading, and sound processing in the brain in preschool children. *npj Science of Learning*, *6*(1), 20. https://doi.org/10.1038/s41539-021-00097-5.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. I. (1997). Training Japanese listeners to identify English/r/and/l: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*(4), 2299-2310.

Brietzke, C., Vinícius, Í., Franco-Alvarenga, P. E., Canestri, R., Goethel, M. F., Santos, L. E. R., ... & Pires, F. O. (2021). Proof-of-concept and test-retest reliability study of psychological and physiological variables of the mental fatigue paradigm. *International Journal of Environmental Research and Public Health*, *18*(18), 9532.

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in second language acquisition*, *33*(3), 433-461. https://doi.org/10.1017/S0272263111000040.

Champely, S. (2020). pwr: Basic Functions for Power Analysis. R package version 1.3-0. https://CRAN.R-project.org/package=pwr.

Christopherson, L. A., & Humes, L. E. (1992). Some psychometric properties of the Test of Basic Auditory Capabilities (TBAC). *Journal of Speech, Language, and Hearing Research*, *35*(4), 929-935.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, *6*(4), 284.

Clayards, M. (2018). Differences in cue weights for speech perception are correlated for individuals within and across contrasts. *J. Acoust. Soc. Am.*, *144*(3), EL172-EL177. https://doi.org/10.1121/1.5052025.

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. https://doi.org/10.1016/j.cognition.2008.04.004.

Coffey, E.B.J., Herholz, S.C., Scala, S. and Zatorre, R.J., 2011, June. Montreal Music History Questionnaire: a tool for the assessment of music-related experience in music cognition

research. In *The Neurosciences and Music IV: Learning and Memory, Conference. Edinburgh, UK*.

Conners, C. K., Epstein, J. N., Angold, A., & Klaric, J. (2003). Continuous performance test performance in a normative epidemiological sample. *Journal of abnormal child psychology*, *31*, 555-562. https://doi.org/10.1023/A:1025457300409.

Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in psychology*, *8*, 1482.

Cowen, A. S., Elfenbein, H. A., Laukka, P., & Keltner, D. (2019). Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*, *74*(6), 698.

Cumming, R., Wilson, A., & Goswami, U. (2015). Basic auditory processing and sensitivity to prosodic structure in children with specific language impairments: A new look at a perceptual hypothesis. *Frontiers in Psychology*, *6*, 972. https://doi.org/10.3389/fpsyg.2015.00972.

Desmeules-Trudel, F. (2018). *Spoken word recognition in native and second language Canadian French: Phonetic detail and representation of vowel nasalization* (Doctoral dissertation, Université d'Ottawa/University of Ottawa).

Desmeules-Trudel, F., & Zamuner, T. S. (2019). Gradient and categorical patterns of spoken-word recognition and processing of phonetic details. *Attention, Perception, & Psychophysics*, *81*, 1654-1672. https://doi.org/10.3758/s13414-019-01693-9.

Destoky, F., Bertels, J., Niesen, M., Wens, V., Vander Ghinst, M., Leybaert, J., ... & Bourguignon, M. (2020). Cortical tracking of speech in noise accounts for reading

strategies in children. *PLoS biology*, *18*(8), e3000840. https://doi.org/10.1371/journal.pbio.3000840.

D'Onofrio, A., Pratt, T., & Van Hofwegen, J. (2019). Compression in the California vowel shift: Tracking generational sound change in California's Central Valley. *Language Variation and Change*, *31*(2), 193-217.

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological bulletin*, *145*(5), 508. https://doi.org/10.1037/bul0000192.

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). In R. Wayland (Ed.), Second language speech learning: Theoretical and empirical progress (pp. 3–83). Cambridge University Press.

Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of phonetics*, *25*(4), 437-470. https://doi.org/10.1006/jpho.1997.0052.

Fox, J., & Weisberg, S. (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Fox-Fuller, J. T., Ngo, J., Pluim, C. F., Kaplan, R. I., Kim, D. H., Anzai, J. A., ... & Quiroz, Y. T. (2022). Initial investigation of test-retest reliability of home-to-home teleneuropsychological assessment in healthy, English-speaking adults. *The Clinical Neuropsychologist*, *36*(8), 2153-2167.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and speech*, *5*(4), 171-189. https://doi.org/10.1177/00238309620050040.

Fuhrmeister, P., & Myers, E. B. (2021). Structural neural correlates of individual differences in categorical perception. *Brain and Language*, *215*, 104919. https://doi.org/10.1016/j.bandl.2021.104919.

Fuhrmeister, P., Phillips, M. C., McCoach, D. B., & Myers, E. B. (2023). Relationships between native and non-native speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001213.

Geller, J., Holmes, A., Schwalje, A., Berger, J. I., Gander, P. E., Choi, I., & McMurray, B. (2021). Validation of the Iowa test of consonant perception. *The Journal of the Acoustical Society of America*, *150*(3), 2131-2153.

Gerrits, E., & Schouten, M. E. (2004). Categorical perception depends on the discrimination task. *Perception & psychophysics*, *66*, 363-376. https://doi.org/10.3758/BF03194885.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and individual differences*, *102*, 74-78.

Giovannone, N., & Theodore, R. M. (2023). Do individual differences in lexical reliance reflect states or traits?. *Cognition*, *232*, 105320.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69-78. https://doi.org/10.1002/wcs.26.

GoldWave Inc. (2015). Goldwave (Version 6.15) [Computer software]. Retreived from: www.goldwave.com.

Golestani, N., & Zatorre, R. J. (2009). Individual differences in the acquisition of second language phonology. *Brain and language*, *109*(2-3), 55-67. https://doi:10.1016/j.bandl.2008.01.005.

González, I., & Déjean, S. (2021). CCA: Canonical Correlation Analysis. R package version 1.2.1. https://CRAN.R-project.org/package=CCA.

Grimaldi, M., Sisinni, B., Gili Fivela, B., Invitto, S., Resta, D., Alku, P., & Brattico, E. (2014). Assimilation of L2 vowels to L1 phonemes governs L2 learning in adulthood: a behavioral and ERP study. *Frontiers in human neuroscience*, *8*, 279. https://doi.org/10.3389/fnhum.2014.00279.

Halperin, J. M., Sharma, V., Greenblatt, E., & Schwartz, S. T. (1991). Assessment of the Continuous Performance Test: Reliability and validity in a nonreferred sample. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*(4), 603.

Hanulíková, A., Dediu, D., Fang, Z., Bašnaková, J., & Huettig, F. (2012). Individual differences in the acquisition of a complex L2 phonology: A training study. *Language Learning*, *62*, 79-109.

Harrell, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer. https://doi.org/10.1007/978-3-319-19425-7.

Hary, J. M., & Massaro, D. W. (1982). Categorical results do not imply categorical perception. *Perception & Psychophysics*, *32*(5), 409-418.

Hattori, K., & Iverson, P. (2009). English/r/-/l/category assimilation by Japanese adults: Individual differences and the link to identification accuracy. *The Journal of the Acoustical Society of America*, *125*(1), 469-479. https://doi.org/10.1121/1.3021295.

Hattori, K., & Iverson, P. (2010). Examination of the relationship between L2 perception and production: an investigation of English/r/-/l/perception and production by adult Japanese speakers. In *Second Language Studies: Acquisition, Learning, Education and Technology*.

Hazan, V., & Rosen, S. (1991). Individual variability in the perception of cues to place contrasts in initial stops. *Perception & Psychophysics*, *49*(2), 187-200.

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. https://doi.org/10.1044/2021_JSLHR-21-00004.

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, *72*(5), 1218-1227. https://doi.org/10.3758/APP.72.5.1218.

Hughes, M. M., Linck, J. A., Bowles, A. R., Koeth, J. T., & Bunting, M. F. (2014). Alternatives to switch-cost scoring in the task-switching paradigm: Their reliability and increased validity. *Behavior research methods*, *46*(3), 702-721. https://doi.org/10.3758/s13428-013-0411-5.

Idemaru, K., Holt, L. L., & Seltman, H. (2012). Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *The Journal of the Acoustical Society of America*, *132*(6), 3950-3964.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47-B57. https://doi.org/10.1016/S0010-0277(02)00198-1.

Jakoby, H., Goldstein, A., & Faust, M. (2011). Electrophysiological correlates of speech perception mechanisms and individual differences in second language attainment. *Psychophysiology*, *48*(11), 1517-1531. https://doi.org/10.1111/j.1469-8986.2011.01227.x.

Joanisse, M. F., Manis, F. R., Keating, P., & Seidenberg, M. S. (2000). Language deficits in dyslexic children: Speech perception, phonology, and morphology. *Journal of experimental child psychology*, *77*(1), 30-60. https://doi:10.1006/jecp.1999.2553.

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. https://doi.org/10.1016/j.bandl.2019.02.004.

Kalaivanan, K., Wong, P. C., Wong, F. C., & Chan, A. H. (2023). Native Language Perceptual Sensitivity Predicts Nonnative Speech Perception Differently in Younger and Older Singaporean Bilinguals. *Journal of Speech, Language, and Hearing Research*, *66*(3), 987-1017. https://doi.org/10.1044/2022_JSLHR-22-00199.

Kang, Y., & Schertz, J. (2021). The influence of perceived L2 sound categories in on-line adaptation and implications for loanword phonology. *Natural Language & Linguistic Theory*, *39*, 555-578. https://doi.org/10.1007/s11049-020-09477-9.

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. https://doi.org/10.1037/xhp0000900.

Kapnoula, E. C., & McMurray, B. (2021). Idiosyncratic use of bottom-up and top-down information leads to differences in speech perception flexibility: Converging evidence from ERPs and eye-tracking. *Brain and Language*, *223*, 105031. https://doi.org/10.1016/j.bandl.2021.105031.

Kapnoula, E. C., & Samuel, A. G. (2023, May 31-June 2). *Examining the links between L1 phoneme categorization and non-native phonetic learning* [Poster presentation]. XVI International Symposium of Psycholinguistics, Vitoria-Gasteiz, Spain.

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. https://doi.org/10.1037/xhp0000410.

Kempe, V., Thoresen, J. C., Kirk, N. W., Schaeffler, F., & Brooks, P. J. (2012). Individual differences in the discrimination of novel speech sounds: effects of sex, temporal processing, musical and cognitive abilities. *PloS one*, *7*(11), e48623. https://doi.org/10.1371/journal.pone.0048623.

Keshavarzi, M., Mandke, K., Macfarlane, A., Parvez, L., Gabrielczyk, F., Wilson, A., ... & Goswami, U. (2022). Decoding of speech information using EEG in children with dyslexia: Less accurate low-frequency representations of speech, not “Noisy” representations. *Brain and Language*, *235*, 105198. https://doi.org/10.1016/j.bandl.2022.105198.

Kim, D., Clayards, M., & Kong, E. J. (2020). Individual differences in perceptual adaptation to unfamiliar phonetic categories. *Journal of Phonetics*, *81*, 100984. https://doi.org/10.1016/j.wocn.2020.100984.

Kogan, V. V., & Mora, J. C. (2022). The effects of individual differences in native perception on discrimination of a novel non-native contrast. *Laboratory Phonology*, *24*(1). https://doi.org/10.16995/labphon.6431.

Kong, E. J., & Edwards, J. (2015, August). Individual differences in L2 learners' perceptual cue weighting patterns. *Proceedings of the International Congress of Phonetic Sciences*.

Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, *59*, 40-57. https://doi.org/10.1016/j.wocn.2016.08.006.

Kong, E. J., & Kang, S. (2022). Individual differences in categorical judgment of L2 stops: A link to proficiency and acoustic cue-weighting. *Language and Speech*, 00238309221108647. https://doi.org/10.1177/00238309221108647.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163.

Kraus, M. S., Gold, J. M., Barch, D. M., Walker, T. M., Chun, C. A., Buchanan, R. W., ... & Keefe, R. S. (2020). The characteristics of cognitive neuroscience tests in a schizophrenia cognition clinical trial: Psychometric properties and correlations with standard measures. *Schizophrenia Research: Cognition*, *19*, 100161.

Krizman, J., & Kraus, N. (2019). Analyzing the FFR: A tutorial for decoding the richness of auditory function. *Hearing research*, *382*, 107779.

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. https://doi.org/10.1098/rstb.2007.2154.

Kwakkel, H., Droop, M., Verhoeven, L., & Segers, E. (2021). The impact of lexical skills and executive functioning on L1 and L2 phonological awareness in bilingual kindergarten. *Learning and Individual Differences*, *88*, 102009. https://doi.org/10.1016/j.lindif.2021.102009.

Lee, S. P. (2016). Computer-detected attention affects foreign language listening but not reading performance. *Perceptual and Motor Skills*, *123*(1), 33-45. https://doi.org/10.1177/0031512516657337

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. https://doi.org/10.1121/1.3506351.

Levy, E. S. (2009). Language experience and consonantal context effects on perceptual assimilation of French vowels by American-English learners of French. *The Journal of the Acoustical Society of America*, *125*(2), 1138-1152. https://doi.org/10.1121/1.3050256.

Li, P., Zhang, F., Tsai, E., & Puls, B. (2013). Language history questionnaire (lhq 2.0): A new dynamic web-based research tool. *Bilingualism: Language and Cognition, 17,* 673-680. https://doi.org/10.1017/S1366728913000606

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.

Linck, J.A. and Weiss, D.J. (2015) Can working memory and inhibitory control predict second language learning in the classroom? *Sage Open Journal*. https://doi.org/10.1177/2158244015607352.

Manis, F. R., McBride-Chang, C., Seidenberg, M. S., Keating, P., Doi, L. M., Munson, B., & Petersen, A. (1997). Are speech perception deficits associated with developmental dyslexia?. *Journal of experimental child psychology*, *66*(2), 211-235.

Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech communication*, *2*(1), 15-35. https://doi.org/10.1016/0167-6393(83)90061-4.

The MathWorks Inc., "MATLAB." Natick, Massachusetts, 2015a.

Mattys, S. L., Carroll, L. M., Li, C. K., & Chan, S. L. (2010). Effects of energetic and informational masking on speech segmentation by native and non-native speakers. *Speech communication*, *52*(11-12), 887-899. https://doi.org/10.1016/j.specom.2010.01.005.

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. https://doi.org/10.1017/S1366728909990022.

McCarthy, C. (2011). The Northern Cities Shift in Chicago. *Journal of English Linguistics*, *39*(2), 166-187.

McMurray, B. (2022). The myth of categorical perception. *The Journal of the Acoustical Society of America*, *152*(6), 3819-3842. https://doi.org/10.1121/10.0016614.

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance, 34*(6), 1609. https://doi.org/10.1037/a0011747.

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. https://doi.org/10.1016/S0010-0277(02)00157-9.

Menzel, U. (2012). CCP: Significance Tests for Canonical Correlation Analysis (CCA). R package version 1.1. https://CRAN.R-project.org/package=CCP.

Miller, J. L. (1994). On the internal structure of phonetic categories: A progress report. *Cognition*, *50*(1-3), 271-285.

Müller, U., Kerns, K. A., & Konkin, K. (2012). Test–retest reliability and practice effects of executive function tasks in preschool children. *The Clinical Neuropsychologist*, *26*(2), 271-287.

Munson, B. (2007). The acoustic correlates of perceived masculinity, perceived femininity, and perceived sexual orientation. *Language and speech*, *50*(1), 125-142.

Munson, B., Johnson, J. M., & Edwards, J. (2012). The role of experience in the perception of phonetic detail in children's speech: a comparison between speech-language pathologists and clinically untrained listeners. *American Journal of Speech-Language Pathology*, *21*(2), 124–139. https://doi.org/ 10.1044/1058-0360(2011/11-0009).

Munson, B., Logerquist, M. K., Kim, H., Martell, A., & Edwards, J. (2021). Does early phonetic differentiation predict later phonetic development? Evidence from a longitudinal study of /ɹ/ development in preschool children. *Journal of Speech, Language, and Hearing Research*, *64*(7), 2417-2437.

Munson, B., Schellinger, S. K., & Edwards, J. (2017). Bias in the perception of phonetic detail in children's speech: A comparison of categorical and continuous rating scales. *Clinical Linguistics & Phonetics*, *31*(1), 56-79. https://doi.org/10.1080/02699206.2016.1233292.

Norrman, G., Bylund, E., & Thierry, G. (2022). Irreversible specialization for speech perception in early international adoptees. *Cerebral Cortex*, *32*(17), 3777-3785. https://doi.org/10.1093/cercor/bhab447.

Ou, J., & Yu, A. C. (2022). Neural correlates of individual differences in speech categorisation: evidence from subcortical, cortical, and behavioural measures. *Language, Cognition and Neuroscience*, *37*(3), 269-284. https://doi.org/10.1080/23273798.2021.1980594.

Ou, J., Yu, A. C., & Xiang, M. (2021). Individual Differences in Categorization Gradience As Predicted by Online Processing of Phonetic Cues During Spoken Word Recognition: Evidence From Eye Movements. *Cognitive Science*, *45*(3), e12948. https://doi.org/10.1111/cogs.12948.

Perkell, J. S., Matthies, M. L., Tiede, M., Lane, H., Zandipour, M., Marrone, N., Stockmann, E., & Guenther, F. H. (2004). The distinctness of speakers' /s/—/ʃ/contrast is related to their auditory discrimination and use of an articulatory saturation effect. *Journal of Speech, Language, and Hearing Research*, *47*(6), 1259–69. https://doi.org/10.1121/1.1787524.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & psychophysics*, *15*(2), 285-290.

Plichta, B., & Preston, D. R. (2005). The /ay/s have it: The perception of /ay/ as a north-south stereotype in United States English. *Acta Linguistica Hafniensia*, *37*(1), 107-130.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language learning*, *64*(4), 878-912.

Power, A. J., Colling, L. J., Mead, N., Barnes, L., & Goswami, U. (2016). Neural encoding of the speech envelope by children with developmental dyslexia. *Brain and Language*, *160*, 1-10. http://dx.doi.org/10.1016/j.bandl.2016.06.006.

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rauber, A. S., Escudero, P., Bion, R. A., & Baptista, B. O. (2005). The interrelation between the perception and production of English vowels by native speakers of Brazilian Portuguese. In *Ninth European Conference on Speech Communication and Technology*.

Saito, K., & Tierney, A. (2022). Domain-general auditory processing as a conceptual and measurement framework for second language speech learning aptitude: A test-retest reliability study. *Studies in Second Language Acquisition*, 1-25.

Schellinger, S. K., Munson, B., & Edwards, J. (2017). Gradient perception of children's productions of /s/ and /θ/: A comparative study of rating methods. *Clinical Linguistics & Phonetics*, *31*(1), 80-103. https://doi.org/10.1080/02699206.2016.1205665.

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of phonetics*, *52*, 183-204.

Schouten, B., Gerrits, E., & Van Hessen, A. (2003). The end of categorical perception as we know it. *Speech communication*, *41*(1), 71-80. https://doi.org/10.1016/S0167-6393(02)00094-8.

Schouten, M. E. H., & Van Hessen, A. J. (1992). Modeling phoneme perception. I: Categorical perception. *The Journal of the Acoustical Society of America*, *92*(4), 1841-1855.

Serniclaes, W., Sprenger-Charolles, L., Carré, R., & Demonet, J. F. (2001). Perceptual discrimination of speech sounds in developmental dyslexia. *Journal of Speech, Language, and Hearing Research*. https://doi.org/10.1044/1092-4388(2001/032)

Serniclaes, W., Ventura, P., Morais, J., & Kolinsky, R. (2005). Categorical perception of speech sounds in illiterate adults. *Cognition*, *98*(2), B35-B44. https://doi.org/10.1016/j.cognition.2005.03.002.

Shen, G., & Froud, K. (2016). Categorical perception of lexical tones by English learners of Mandarin Chinese. *The Journal of the Acoustical Society of America*, *140*(6), 4396-4403. https://doi.org/10.1017/S136672891800038X.

Sherry, A., & Henson, R. K. (2005). Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of personality assessment*, *84*(1), 37-48.

Silbert, N. H., Smith, B. K., Jackson, S. R., Campbell, S. G., Hughes, M. M., & Tare, M. (2015). Non-native phonemic discrimination, phonological short term memory, and word learning. *Journal of Phonetics*, *50*, 99-119. https://doi.org/10.1016/j.wocn.2015.03.001.

Silveira, R. (2011). Pronunciation instruction and syllabic-pattern discrimination. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, *27*, 21-36. https://doi.org/10.1590/S0102-44502011000100002.

Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter?. *Psychological science*, *17*(8), 675-681. https://doi.org/10.1111/j.1467-9280.2006.01765.x

Souza, P., Wright, R., Gallun, F., & Reinhart, P. (2018). Reliability and repeatability of the speech cue profile. *Journal of Speech, Language, and Hearing Research*, *61*(8), 2126-2137.

Spreng, R. N., & Turner, G. R. (2019). The shifting architecture of cognition and brain function in older adulthood. *Perspectives on Psychological Science*, *14*(4), 523-542. https://doi.org/10.1177/1745691619827511.

Sprouse, J. (2007). Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics*, *1*, 123-134. https://doi.org/10.5964/bioling.8597.

Strand, J., Cooperman, A., Rowe, J., & Simenstad, A. (2014). Individual differences in susceptibility to the McGurk effect: Links with lipreading and detecting audiovisual incongruity. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2322-2331. https://doi.org/10.1044/2014_JSLHR-H-14-0059.

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & psychophysics*, *36*(2), 131-145. https://doi.org/10.3758/BF03202673.

Strange, W., & Shafer, V. L. (2008). Speech perception in second language learners. In J. G. H. Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp.153-191). Amsterdam: John Benjamins.

Surprenant, A. M., & Watson, C. S. (2001). Individual differences in the processing of speech and nonspeech sounds by normal-hearing listeners. *The Journal of the Acoustical Society of America*, *110*(4), 2085-2095. https://doi.org/10.1121/1.1404973.

Tecoulesco, L., Skoe, E., & Naigles, L. R. (2020). Phonetic discrimination mediates the relationship between auditory brainstem response stability and syntactic performance. *Brain and Language*, *208*, 104810. https://doi.org/10.1016/j.bandl.2020.104810.

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. https://doi.org/10.1177/0956797610384142.

UCLA: Statistical Consulting Group. *Canonical Correlation Analysis: R Data Analysis Examples.* UCLA Statistical Methods and Data Analytics. https://stats.oarc.ucla.edu/r/dae/canonical-correlation-analysis/.

Wang, X., & Humes, L. E. (2008). Classification and cue weighting of multidimensional stimuli with speech-like cues for young normal-hearing and elderly hearing-impaired listeners. *Ear and hearing*, *29*(5), 725.

Wechsler, D. (2008). Wechsler adult intelligence scale–Fourth Edition (WAIS–IV). *San Antonio, TX: NCS Pearson*, *22*(498), 1.

Werker, J. F., & Tees, R. C. (1987). Speech perception in severely disabled and average reading children. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *41*(1), 48. https://doi.org/doi: 10.1037/h0084150.

White-Schwoch, T., Woodruff Carr, K., Thompson, E. C., Anderson, S., Nicol, T., Bradlow, A. R., ... & Kraus, N. (2015). Auditory processing in noise: A preschool biomarker for literacy. *PLoS biology*, *13*(7), e1002196. https://doi.org/10.1371/journal.pbio.1002196.

Won, J. H., Tremblay, K., Clinard, C. G., Wright, R. A., Sagi, E., & Svirsky, M. (2016). The neural encoding of formant frequencies contributing to vowel identification in normal-hearing listeners. *The Journal of the Acoustical Society of America*, *139*(1), 1-11. https://doi.org/10.1121/1.4931909.

Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., ... & Reed, B. (2011). Improving digit span assessment of short-term verbal memory. *Journal of clinical and experimental neuropsychology*, *33*(1), 101-111.

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. https://doi.org/10.3758/s13414-017-1361-2.

Yu, A. C., & Lee, H. (2014). The stability of perceptual compensation for coarticulation within and across individuals: A cross-validation study. *The Journal of the Acoustical Society of America*, *136*(1), 382-388. https://doi.org/10.1121/1.4883380.

Zhang, J., & Mueller, S. T. (2005). A note on ROC analysis and non-parametric estimate of sensitivity. *Psychometrika*, *70*, 203-212. https://doi.org/10.1007/s11336-003-1119-8.