Attention, Perception, & Psychophysics (2020) 82:2049–2065
https://doi.org/10.3758/s13414-019-01925-y

Desirable and undesirable difficulties: Influences of variability, training schedule, and aptitude on nonnative phonetic learning

Pamela Fuhrmeister[1] · Emily B. Myers[2]

Published online: 22 January 2020

Abstract

Adult listeners often struggle to learn to distinguish speech sounds not present in their native language. High-variability training sets (i.e., stimuli produced by multiple talkers or stimuli that occur in diverse phonological contexts) often result in better retention of the learned information, as well as increased generalization to new instances. However, high-variability training is also more challenging, and not every listener can take advantage of this kind of training. An open question is how variability should be introduced to the learner in order to capitalize on the benefits of such training without derailing the training process. The current study manipulated phonological variability as native English speakers learned a difficult nonnative (Hindi) contrast by presenting the nonnative contrast in the context of two different vowels (/i/ and /u/). In a between-subjects design, variability was manipulated during training and during test. Participants were trained in the evening hours and returned the next morning for reassessment to test for retention of the speech sounds. We found that blocked training was superior to interleaved training for both learning and retention, but for learners in the interleaved training group, higher pretraining aptitude predicted better identification performance. Further, pretraining discrimination aptitude positively predicted changes in phonetic discrimination after a period of off-line consolidation, regardless of the training manipulation. These findings add to a growing literature suggesting that variability may come at a cost in phonetic learning and that aptitude can affect both learning and retention of nonnative speech sounds.

Keywords Variability · Nonnative phonetic learning · Consolidation · Individual differences

Acquiring perceptual sensitivity to nonnative speech contrasts is challenging for many adult learners. A great deal of work has focused on different types of training to optimize learning of difficult speech sounds (e.g., Lim & Holt, 2011; Logan, Lively, & Pisoni, 1991; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002; Vlahou, Protopapas, & Seitz, 2012), but even so, success is quite variable among individuals (Golestani & Zatorre, 2004; Myers & Swan, 2012; Yi, Maddox, Mumford, & Chandrasekaran, 2014). Studies have found positive effects of exposing learners to variability during training (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Lively, Logan, & Pisoni, 1993; Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994; Logan et al., 1991), but others have found that variability can come at a cost for learning or retention (Fuhrmeister & Myers, 2017; Perrachione, Lee, Ha, & Wong, 2011; Sadakata & McQueen, 2014). The current study investigates how the introduction of variability in both training and testing environments relates to learning and retention of nonnative speech sounds.

The original version of this article was revised: Due to a production error, some IPA symbols were not included.

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13414-019-01925-y) contains supplementary material, which is available to authorized users.

✉ Pamela Fuhrmeister
pamela.fuhrmeister@uconn.edu

[1] University of Connecticut, Storrs, CT, USA

[2] Department of Speech, Language, and Hearing Sciences, University of Connecticut, 2 Alethia Dr, Storrs, CT 06269, USA

Variability in learning

Several accounts, both within and outside the speech learning literature, suggest that exposure to variability during learning, especially perceptual learning, should benefit both learning and retention of new information, as well as generalization of that knowledge to novel contexts (Battig, 1972; Bradlow et al., 1999; Bradlow & Bent, 2008; Bradlow et al., 1997;

Lively et al., 1993; Lively et al., 1994; Logan et al., 1991; Shea & Morgan, 1979). In the case of nonnative speech sound learning, variability is typically introduced by presenting speech sounds spoken by a variety of talkers or occurring in multiple phonological contexts. For instance, a difficult nonnative speech sound contrast (such as the /ɹ/–/l/ contrast for Japanese learners of English), might be adjacent to any number of vowels ("rock," "leaf"), and in different syllable positions in the word (e.g., Logan et al., 1991). Because speech gestures for neighboring sounds tend to overlap (are coarticulated), these diverse phonological environments also pervade the to-be-learned speech sound, resulting in more acoustic variability in the set of novel speech tokens. However, the notion of variability in learning can be understood in two senses: The first is the aforementioned variability in the stimulus set (see Bradlow et al., 1999, Bradlow et al., 1997; Lively et al., 1993; Lively et al., 1994; Logan et al., 1991); second, training paradigms differ in the degree of trial-to-trial variability, even while keeping the overall variability in the stimulus set constant. Consider for instance, a paradigm using phonetic tokens spoken by four different talkers. A blocked training schedule, in which one talker is presented per block, minimizes trial-to-trial variability, but interleaving tokens spoken by all talkers results in high trial-to-trial variability. As we discuss below, these two types of variability—that is, variability in the stimulus set and in the training schedule—have consequences for learning.

The idea that variability enhances learning has a long history: the *contextual interference effect* is a domain-general learning theory that posits that random practice is superior to blocked practice when learning a skill, particularly when it comes to retention and generalization of that skill (e.g., Battig, 1972; Shea & Morgan, 1979; see Magill & Hall, 1990, for review). In random practice, different types of trials are interleaved, which creates a situation of high contextual interference. In other words, the variable trial types interfere with each other during learning, resulting in more difficult learning conditions. Conversely, blocking different trial types during practice creates low contextual interference conditions. Conditions of high contextual interference typically result in poorer performance during practice but superior performance when retention or generalization of the skill is tested (e.g., Li & Wright, 2000; Magill & Hall, 1990; Shea & Morgan, 1979). For example, Shea and Morgan (1979) found that participants who learned a motor task were more successful in generalizing that skill to a new sequence if practice trials were presented in random order rather than blocked order, especially when the new sequence was more complex. In another study examining learning of a motor task, Li and Wright (2000) observed that participants who learned key pressing sequences with random practice showed better retention of the sequences than groups who had blocked practice. While these advantages have been predominantly explored in motor learning studies (e.g., Hall, Domingues, & Cavazos, 1994; Li & Wright, 2000; Shea & Morgan, 1979), this concept has also been extended to foreign language word learning (Schneider, Healy, & Bourne, 1998, 2002). For instance, in Schneider et al. (2002), participants learned English–French vocabulary pairs whose presentation during learning was either grouped (blocked) by category or mixed. Learners who had mixed practice remembered more word pairs than those who had blocked practice when tested a week later.

Several studies have probed the efficacy of random, or interleaved, practice in phonetic learning, as well. There are many reasons to believe that exposure to variability, specifically in phonetic training, may be beneficial to the learner. Due to the inherent variability in the speech signal (e.g., Hillenbrand, Getty, Clark, & Wheeler, 1995; Peterson & Barney, 1952), learners may not have enough information to discover the boundaries of a new phonetic category when only exposed to limited instances of the speech sounds (e.g., speech sounds produced by a single talker or presented in a single phonological context). Indeed, variability may encourage attention to invariant features of the stimuli, which will aid in category acquisition (see Apfelbaum & McMurray, 2011; Galle, Apfelbaum, & McMurray, 2015, for evidence in infant word learning). Other studies have demonstrated that individuals show improved perception and generalization after exposure to speech sounds that are spoken by a variety of talkers or presented in various word positions or phonological contexts (i.e., variability in the stimulus set) when learning nonnative speech contrasts (Bradlow et al., 1999; Bradlow et al., 1997; Lively et al., 1993; Lively et al., 1994; Logan et al., 1991) or learning a nonnative accent (Bradlow & Bent, 2008). Notably in several of these studies (Bradlow et al., 1999; Bradlow et al., 1997; Lively et al., 1993; Lively et al., 1994; Logan et al., 1991), phonological context was variable from trial to trial, but the talker was blocked, making it difficult to conclude whether interleaved or blocked practice is most effective. Similar to studies testing the contextual interference effect, the strongest benefits seen as a result of these high-variability training paradigms are typically in generalization or transfer of knowledge to novel talkers or phonological contexts (Bradlow & Bent, 2008; Lively et al., 1993). Studies by Bradlow and colleagues (Bradlow et al., 1999; Bradlow et al., 1997) have even observed that phonetic training in *perception* generalized to speech *production* (Bradlow et al., 1999; Bradlow et al., 1997). High-variability training also promotes long-term retention of speech sounds (Lively et al., 1994) and even long-term retention of improvements in production that resulted from high-variability perceptual training (Bradlow et al., 1999).

In addition to facilitating the discovery of the perceptual space that a speech category occupies, variability presented in training (specifically trial-by-trial variability) may influence whether an individual engages in optimal learning strategies.

The dual-systems model of speech category learning specifies the learning systems involved with this process (Chandrasekaran, Koslov, & Maddox, 2014; Chandrasekaran, Yi, & Maddox, 2014). According to this account, two dissociable systems underlie speech category learning: the *reflective* system is a rule-based learning system that makes use of verbalizable rules and explicit, detailed feedback during training in order to generate, test, and modify hypotheses about the categories being learned. The *reflexive* system does not take advantage of verbalizable rules; rather, it learns via activation of a reward system as a result of a motor action in response to a stimulus. Importantly, reliance on the reflexive system is optimal for speech category learning, as the speech signal does not easily lend itself to categorization based on verbalizable rules (Chandrasekaran, Koslov, et al., 2014). In addition, information must often be integrated across multiple dimensions of the speech signal in order to accurately categorize speech sounds, and participants who use the reflexive system or implicit learning strategies typically demonstrate superior speech category learning of multidimensional stimuli (Chandrasekaran, Koslov, et al., 2014; Yi et al., 2014; see also Wade & Holt, 2005, for auditory category learning). Furthermore, Chandrasekaran, Koslov, et al. (2014) maintain that variability in training encourages use of the optimal reflexive system because the trial-by-trial variability does not allow a learner to predict certain features of the speech signal on each presentation, nor does it allow learners to easily develop verbalizable hypotheses about the speech categories.

Despite the many reported advantages of variable or random practice for learning a new skill, some findings challenge whether variability is universally beneficial. First, Wulf and Shea (2002) argue that the benefits of random practice may not hold if the processing demands are too high for the learner. Furthermore, they caution against extending findings from the motor learning studies that have explored this effect because the majority of those studies have tested high contextual interference conditions when unchallenging skills were learned. They argue that learners may not improve on skills that are more complex if the processing demands are too high during learning. It is reasonable to believe that learning certain nonnative speech sound contrasts is a more complex skill, especially for speech sounds that are close in perceptual space to native-language phonemes, as these sounds are extremely difficult for adult learners (e.g., Best, McRoberts, & Goodell, 2001). Indeed, some studies in the speech domain have found limitations of variability in training or task performance. For instance, Mullennix, Pisoni, and Martin (1989) found that participants who performed spoken word recognition tasks were not as accurate and had slower reaction time latencies when words were presented in a multitalker condition as opposed to a single-talker condition.

Another possibility is that variability interferes with memory consolidation processes and therefore diminishes retention of learned phonetic information (Fuhrmeister & Myers, 2017; see also Brown & Gaskell, 2014, for evidence in word learning). Several studies suggest that learners are better able to consolidate *strongly* learned information, as opposed to *weakly* learned information (Ebbinghaus, 1885; Hauptmann, Reinhart, Brandt, & Karni, 2005; Shibata et al., 2017; Tucker & Fishbein, 2008; but see Drosopoulos, Schulze, Fischer, & Born, 2007). If the introduction of variability, especially trial-by-trial variability resulting from random/interleaved practice, leads to less-stable learning, we would predict that consolidation of this information may be weaker for participants exposed to high-variability training, leading to poorer retention in phonetic learning. In other words, variability may affect retention because of how it affects learning.

Supervised versus unsupervised variability

Recent findings suggest that exposure to variability without explicit feedback can affect nonnative speech sound learning. A study from our group found that even when variability was introduced at test rather than training, nonnative learning was destabilized (Fuhrmeister & Myers, 2017). In this study, all participants were trained on a nonnative dental/retroflex stop contrast in one vowel context (e.g., "ɖu" and "d̪u"). Half of the participants were tested with an additional, untrained vowel context in a pretest, immediate posttest, and a delayed posttest. These participants demonstrated poorer performance overall and no improvement after a period of sleep as compared with a group who was only exposed to the sounds in one vowel context throughout training and assessments. This suggests that even minimal exposure to phonological variability during test compromised learning and retention of new phonetic information. However, one potential explanation for this finding is that variability was introduced during testing only; therefore, learners were not explicitly trained with feedback on the speech sounds in both contexts. It is possible that training participants and providing feedback on the sounds presented in both vowel contexts could "rescue" participants from the detrimental effects of introducing variability during assessments. This idea is consistent with category learning studies examining the effects of supervised and unsupervised learning (i.e., with or without feedback). Several studies have found that trial-by-trial feedback in learning is helpful or even necessary when learning visual or speech categories that differ on multiple dimensions (e.g., Ashby, Queller, & Berretty, 1999; Goudbeek, Cutler, & Smits, 2008). In general, learners must attend to multiple acoustic cues to differentiate speech sounds. For example, for the Hindi dental-retroflex contrast used in Fuhrmeister and Myers (2017), participants likely had to learn to attend to both burst frequency and formant transitions in order to differentiate retroflex from dental sounds (Stevens & Blumstein, 1975). The addition of the second vowel context for some learners may have increased the within-category variability of the categories,

which can lead to poorer performance in unsupervised learning tasks (Ell & Ashby, 2012). Therefore, the "unsupervised" variability present in Fuhrmeister and Myers (2017) may have been confusing to this group of learners, and an unanswered question is whether feedback might eliminate the detrimental effects of this variability.

Individual predictors in phonetic learning

Learners vary greatly in their ability to acquire perceptual sensitivity to novel speech sound contrasts (e.g., Golestani & Zatorre, 2004; Myers & Swan, 2012; Yi et al., 2014). Studies by Perrachione and colleagues (2011) and Sadakata and McQueen (2014) explored the relationship between pretraining aptitude and high-variability training in a nonnative, lexical tone learning task. In the study by Perrachione et al. (2011), aptitude was measured by a pretraining pitch contour perception task, and Sadakata and McQueen (2014) determined aptitude by pretraining and posttraining identification of the tonal contrast of interest. In these studies, only high-aptitude learners benefitted from high-variability training. In fact, the study by Perrachione and colleagues showed that the trial-by-trial nature of the high-variability training paradigm was what was detrimental to learning for individuals with poorer perceptual abilities (i.e., those with lower pretraining aptitude). Antoniou and Wong (2015) found that poor perceivers are more affected by increases in cognitive load, which may explain why poor perceivers are more adversely affected by high-variability training. Taken together, it appears that the advantages of high-variability training in phonetic learning may be limited to individuals with stronger perceptual abilities (i.e., individuals who can detect subtle acoustic differences in the speech categories prior to training). The increased difficulty of variability in training may have negative consequences for learning and retention, depending on the aptitude of the individual learner.

Current study

In the current study, we trained all participants to learn the Hindi dental and retroflex voiced stop consonants (/d̪/ and /ɖ/) in word-initial position in consonant–vowel–consonant (CVC) contexts with two different vowels (/i/ and /u/). For half of the participants, training was blocked (one vowel context at a time), and for the other half, training trials with both vowel contexts were interleaved. As in Fuhrmeister and Myers (2017), half of the participants heard both vowel contexts at test and half heard only one vowel context. This training and testing structure resulted in four groups in a two-by-two design (one-vowel test: blocked training; one-vowel test: interleaved training; two-vowel test: blocked training; and two-vowel test: interleaved training), which are further described below.

The goals of the current study are to better understand the consequences of variability in phonetic learning and retention. Specifically, we test the hypothesis that retention of newly learned phonetic categories (after an approximately 12-hour, overnight interval) will be disrupted by variability present in training and testing. Further, we hypothesize that individuals who receive blocked training (vs. interleaved) will show the maximum benefit of learning and retention, supporting the notion that strongly encoded information is better retained. However, a finding that interleaved training better supports retention would be consistent with the contextual interference effect, in which more difficult training conditions result in better retention of information. We additionally assess whether differences in individual aptitude predict next-day performance on phonetic discrimination and identification and whether this relationship changes based on the training condition that participants receive (blocked vs. interleaved) or the number of vowels learners are exposed to in test (one vs. two). If more strongly learned information is more easily consolidated, we predict that high-aptitude learners will demonstrate better consolidation of the material (as indicated by a stronger relationship between aptitude and next-day task performance than same-day task performance). Based on previous literature (Perrachione et al., 2011; Sadakata & McQueen, 2014), we expect to see this pattern of results among participants who received interleaved training; however, we may see it among all participants, as higher pretraining discrimination abilities likely allow participants to learn the contrast more robustly, regardless of the training they received. Last, we test the hypothesis that providing learners with explicit feedback on all tokens heard during training and testing will alleviate some of the unfavorable effects of unsupervised variability seen in Fuhrmeister and Myers (2017). If the unsupervised variability during test continues to be detrimental to learning even when listeners also receive feedback on both vowel contexts during training, this would support a model in which passive exposure to variability destabilizes learning, even when explicit feedback on those tokens is provided during training. In contrast, if no differences emerge between participants tested on one versus two vowels, this supports the notion that providing contextual cues in the form of feedback allows listeners to recover from the variability penalty introduced during test.

Method

Participants

A total of 166 participants were recruited through the University of Connecticut Psychology Department participant pool. Thirty participants were excluded from the analyses due

to failure to complete both sessions of the experiment (19 participants), noncompliance with experimental tasks (6), and computer or experimenter errors (5). The remaining 136 participants (83 female, 49 male; age range: 18–22 years; age and gender were not collected for four participants due to experimenter error) are included in the analyses described below (see Table 1). The study was advertised to monolingual, native speakers of North American English only, and participants reported no history of speech or language disorders, typical hearing and vision, and no exposure to foreign languages or accented speech from parents or primary caregivers during childhood. Participants received course credit for their participation and gave informed consent according to the University of Connecticut Institutional Review Board procedures.

Stimuli

Auditory stimuli were recorded by a female, native speaker of Hindi in a soundproof booth using a Roland R-05 digital voice recorder. Five acoustically distinct exemplars each of /d̪ug/, /d̪ig/, /d̪ug/, and /d̪ig/ were recorded, and stimuli were scaled to a mean amplitude of 65 dB sound pressure level in Praat (Boersma & Weenink, 2018). Visual stimuli consisted of two distinct "Fribbles" (stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, https://www.tarrlab.org/) and two distinct images from the Novel Object and Unusual Name (NOUN) database (Horst & Hout, 2016). All experimental stimuli were presented using Open Sesame (Mathôt, Schreij, & Theeuwes, 2012), and auditory stimuli were played over headphones (SONY MDF-7606) at a comfortable listening level that participants could adjust. Participants indicated their responses by pressing the appropriate key on a keyboard.

Procedure

Participants made two visits to the lab. The first session took place in the evening hours between 5:00 p.m. and 9:00 p.m., and the second session occurred the following morning between 8:00 a.m. and 10:00 a.m. Participants were randomly assigned to one of four groups according to a 2 × 2 design (see Fig. 1). The test manipulation introduced either one or two vowels during the assessments (see one-vowel test, two-vowel test), and the training manipulation introduced either blocked training (one vowel context at a time) or interleaved training (both vowel contexts randomly interleaved).

Baseline Participants completed a baseline assessment of their ability to discriminate the Hindi dental and retroflex speech sounds by means of an AX discrimination task. In this task, participants heard two tokens separated by a 1-second inter-stimulus interval and were asked to indicate whether they began with the same or different speech sounds. Half of the trials corresponded to two examples of the same speech sound (dental–dental or retroflex–retroflex) and half corresponded to two different sounds (retroflex–dental or dental–retroflex). Importantly, the two audio files in a pair were always distinct recordings to discourage participants from using low-level details of the signal to discriminate tokens. The one-vowel groups received 64 trials in the /u/ vowel context, and the two-vowel groups heard 128 trials: 64 in the /u/ context and 64 in the /i/ context. For the two-vowel groups, /u/ and /i/ trials during test were randomized.

Training Following the baseline assessment, participants completed a training task to learn the dental versus retroflex contrast. Participants were told that they would learn four new words (/d̪ug/, /d̪ug/, /d̪ig/, and /d̪ig/). Each novel word was paired with a novel object, and participants were first familiarized with each word–object pairing by seeing the novel object on the screen and hearing a corresponding auditory stimulus five times in a row for each pairing. After familiarization, participants were presented with either two visual objects at a time (blocked training groups) or four objects (interleaved training groups) on the screen and heard one auditory token per trial. Participants were asked to indicate which visual object belonged to the nonword they heard by pressing the corresponding button on the keyboard, which was displayed directly under the visual object on the screen throughout the duration of the task. All participants received

Table 1 Breakdown of number of participants, mean, and standard deviation (*SD*) of the time between sessions in hours, gender, and age by group

Group	N	Time between	sd	Female	Age	sd
One vowel test: blocked training	35	14.41	1.40	19	18.58	0.92
One vowel test: interleaved training	33	14.86	1.00	19	19.12	1.04
Two vowel test: blocked training	34	14.25	2.60	23	19.09	1.08
Two vowel test: interleaved training	34	14.75	1.30	22	18.71	0.97

Demographic data was not collected for four participants in the one-vowel test: blocked training group due to experimenter error, so data reported for that group is from the 31 participants whose demographic data were collected. In addition, one participant in the one-vowel test: interleaved training group did not wish to provide his or her age, so summarized age data in that group includes the remaining 32 participants

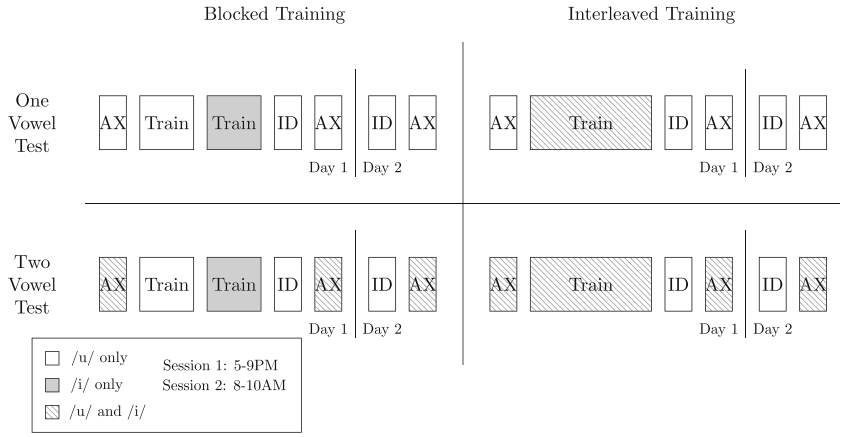

Fig. 1 Schematic of training and testing manipulations for all four groups in the 2 × 2 design (one-vowel test: blocked training, one-vowel test: interleaved training, two-vowel test: blocked training, and two-vowel test: interleaved training). Train = training task, ID = identification test, and AX = (AX) discrimination test. All groups were trained on the Hindi sounds in two vowel contexts: /u/ and /i/. Blocked training groups were trained on the sounds in a blocked manner, /u/ followed by /i/, and interleaved training groups were trained on the sounds in both vowel contexts at once, and the trials were randomized. One-vowel test groups were only tested on the sounds presented in the vowel context /u/, whereas the two-vowel test groups were tested on the sounds in both the /u/ and /i/ vowel contexts in the AX discrimination assessments in random order

600 trials total in training (300 per vowel context). The blocked groups heard 300 trials in the /u/ vowel context followed by 300 in the /i/ context, and the interleaved groups also completed 300 trials of each vowel context, but these trials were interleaved. Visual feedback was presented after each trial (e.g., "Correct!" or "Incorrect").

Posttraining assessments Immediately after training, all participants completed an identification assessment consisting of 50 trials of the Hindi sounds presented in the /u/ vowel context only. This test was identical to the training block, except that no feedback was given and interleaved training groups only saw two objects on the screen at once. Importantly, this assessment was identical for every group. Next, participants completed an additional AX discrimination assessment according to the condition they were assigned to. The following morning, participants returned for reassessment to measure their retention of the Hindi sounds. Reassessments included identification and AX discrimination tests.

Analysis approach Mixed effects logistic regression models were used for analysis of identification tasks in training and assessments, and linear mixed effects models were used for analysis of discrimination data. Mixed effects models were conducted using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in R (R Development Core Team, 2008), and *p* values for linear mixed effects models were estimated using the afex package (Singmann, Bolker, Westfall & Aust, 2019). For mixed effects models that included by-item random effects, items consisted of the individual sound files for each nonword that participants learned (/d̪ug/, /d̪ug/, /d̪ig/, and /d̪ig/). A backwards stepping procedure was used in the analyses of training and identification to determine the random effects structure best justified by the data (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). For all analyses of the identification task, we assumed that participants who scored a mean percent accuracy that was significantly below chance as indicated by a binomial test could indeed identify the speech sounds but simply switched the labels of the sounds. The binomial test indicated that the probability of obtaining an accuracy score of less than .38 was significantly below chance ($p < .05$). We recoded responses from those participants at the appropriate posttest to indicate that the labels were switched (i.e., 0 was recoded as 1 and 1 was recoded as 0).[1] This manipulation affected a small proportion of the total participants (see Identification Performance, below, for specific details).

Results

Training

We first wanted to determine whether training manipulations (blocked vs. interleaved) or the number of vowels presented in the discrimination tests (one vs. two) would have an effect on performance during training. The contextual interference effect would predict that interleaved training should be more difficult; thus, we expect a main effect of training, such that interleaved groups have lower accuracy than blocked groups during training. Based on previous findings from Fuhrmeister and Myers (2017), we predict that exposure to the second vowel context in the discrimination *pretest* in the two vowel groups will hinder performance during training.

[1] Data files and analysis scripts for this project can be found at https://osf.io/ujm4f/.

To test for group differences in training performance, we ran a mixed effects logistic regression model using the log odds of selecting a correct response as the dependent variable. Fixed effects included vowel context (/u/ or /i/), training condition (blocked or interleaved), and number of vowels in the discrimination assessments (one or two). The random effects structure included by-subject intercepts and slopes for vowel context and by-item intercepts and slopes for training condition, number of discrimination vowels, and their interactions (the maximal random effects structure). Factors were deviation coded (−.5, .5) to test for main effects. Factor levels were coded as follows: vowel context (/i/ = −.5, /u/ = .5), training condition (blocked = −.5, interleaved = .5), number of vowels in the discrimination task (one = −.5, two = .5).

This analysis revealed a main effect of training condition, $\beta = -0.51$, $SE = 0.13$, $z = -4.05$, $p < .001$, such that the blocked training group significantly outperformed the interleaved training group. Note that the interleaved group saw four alternatives at a time, whereas the blocked group saw two alternatives at a time. We discuss this issue further below. In addition to the main effect of training type, we observed a main effect of vowel context, $\beta = 0.41$, $SE = 0.11$, $z = 3.60$, $p < .001$, with participants showing significantly better accuracy with /u/ vowel trials than /i/ vowel trials. We observed no main effect of the number of vowels in the discrimination task, nor any interactions (see Table 2, Fig. 2).

Identification performance

To measure learning and retention of the phonetic contrast in the identification task, we carried out a mixed effects logistic regression model. Recall that only the /u/ vowel context was tested in the identification tests, so this task (a two-alternative forced choice [2AFC]) was identical for all groups. Data were recoded for eight participants (who switched the labels of the sounds) on Day 1 (three participants from the one-vowel test: blocked training group, three from the two-vowel test: blocked training group, and two from the one-vowel test: interleaved training group) and for 10 participants on Day 2 (four in the one-vowel test: blocked training group, five in the two-vowel test: blocked training group, and one in the one-vowel test: interleaved training group). The dependent variable of the model was the log odds of selecting a correct response, and fixed effects included training (blocked or interleaved), the number of vowels participants were exposed to during the discrimination test (one vs. two), and time (Day 1 or Day 2). In the final model, by-participant random intercepts and slopes for time point and by-item random intercepts were included, and correlation parameters were set to zero. To allow models to converge, we used the optimizer "bobyqa" in the glmer control options and increased the number of iterations to 200,000. Factors were deviation coded (−.5, .5) to test for main effects, and factor levels were coded as follows: training condition (blocked = −.5, interleaved = .5), number of vowels in the discrimination test (one = −.5, two = .5), and time (Day 1 = −.5, Day 2 = .5).

Table 2 Descriptive statistics for performance on training task by group and vowel context

Group	Vowel context	N	mean	sd
One vowel test: blocked training	/i/	35	0.68	0.13
One vowel test: interleaved training	/i/	33	0.59	0.13
Two vowel test: blocked training	/i/	34	0.67	0.11
Two vowel test: interleaved training	/i/	34	0.57	0.12
One vowel test: blocked training	/u/	35	0.75	0.17
One vowel test: interleaved training	/u/	33	0.67	0.17
Two vowel test: blocked training	/u/	34	0.73	0.14
Two vowel test: interleaved training	/u/	34	0.62	0.14

This analysis revealed a significant main effect of training, $\beta = -0.47$, $SE = 0.23$, $z = -2.01$, $p = .04$, with better performance for participants who had received blocked compared with interleaved training, a significant main effect of time, $\beta = 0.27$, $SE = 0.07$, $z = 3.65$, $p < .001$, with better performance on Day 2, no main effect of number of vowels presented in the discrimination test, and no interactions (see Table 3, Fig. 3). Thus, it appears that blocked training is not necessarily advantageous for *retention* of phonetic information, but rather the blocked training groups maintain the benefits conferred during initial learning of the contrast.

Discrimination performance

Discrimination: Learning In order to measure group differences in how well participants were able to learn the phonetic contrast, we carried out a linear mixed effects model that included data from the pretest and immediate posttest for the /u/ trials only (the common vowel context among groups). We first calculated *d* prime (*d'*) scores to account for response bias (MacMillan & Creelman, 2005), and these served as the dependent measure in the model. Fixed effects included training (blocked vs. interleaved), number of vowels in the discrimination test (one or two), and time (pretest vs. immediate posttest). By-participant random intercepts were included, as well.[2] Factors were deviation coded (−.5, .5) to test for main effects: training (blocked = −.5, interleaved = .5), number of vowels in the discrimination test (one = −.5, two = .5), and time (pretest = −.5, immediate posttest = .5). The analysis indicated a significant main effect of time, $\beta = 0.52$, $SE = 0.08$, $t = 6.91$, $p < .001$, suggesting that, as predicted, participants improved after training. In addition, we observed a nonsignificant trend towards an effect of training, $\beta = -0.22$,

[2] Note that because we averaged over trials in this analysis (and all analyses of the discrimination task), it was not possible to include random slopes because there was not enough data to estimate them.

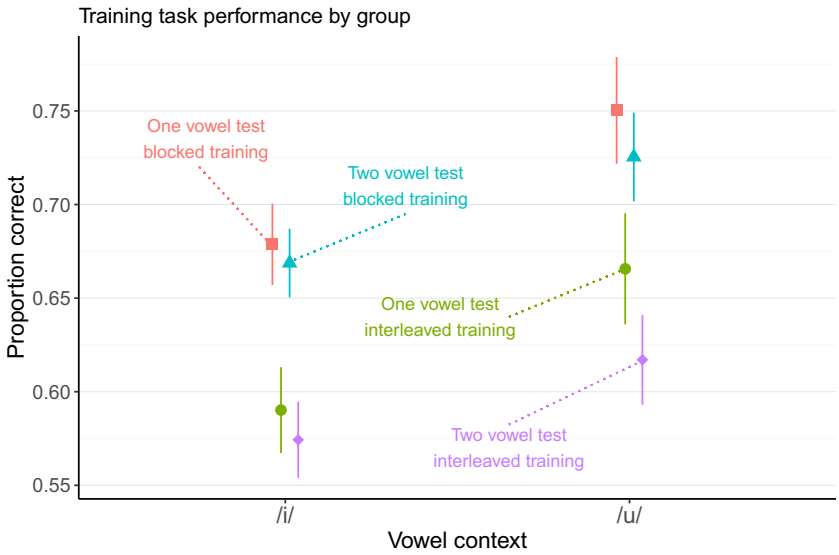

Fig. 2 Proportion correct responses during training for each group and each vowel context. The blocked training groups performed significantly better than the interleaved training groups, and overall, performance was significantly better on /u/ vowel trials than /i/ trials. Error bars denote standard error of the mean

SE = 0.12, *t* = −1.84, *p* = .07, and a nonsignificant trend towards an interaction between the number of vowel contexts in the discrimination test and time, β = −0.27, *SE* = 0.15, *t* = −1.79, *p* = .08. No other interactions were significant or approaching significance (see Table 4, Fig. 4).

Discrimination: Retention To test group differences in the retention of the speech sounds, an additional linear mixed effects model was carried out. This model was identical to the one described above to measure learning, except this model included data from the immediate posttest and the next-day posttest to measure how well participants retained the information from one day to the next. Factors were again deviation coded (−.5, .5) just as in the previous analysis, except the factor time was coded in this analysis as immediate posttest = −.5, next-day posttest = .5. No significant effects of time, training condition, or number of vowels presented in the tests were found, and neither were any significant interactions. However, a nonsignificant trend towards an effect of training was found, β = −0.33, *SE* = 0.18, *t* = −1.86, *p* = .06 (see Table 5, Fig. 4).

Table 3 Descriptive statistics for performance on identification tests by group and time point

Group	Time	N	mean	sd
One vowel test: blocked training	Day 1	35	0.78	0.16
One vowel test: interleaved training	Day 1	33	0.73	0.19
Two vowel test: blocked training	Day 1	34	0.77	0.19
Two vowel test: interleaved training	Day 1	34	0.67	0.16
One vowel test: blocked training	Day 2	35	0.81	0.18
One vowel test: interleaved training	Day 2	33	0.76	0.17
Two vowel test: blocked training	Day 2	34	0.77	0.18
Two vowel test: interleaved training	Day 2	34	0.70	0.19

Discrimination: Group analysis Although the main analysis for the discrimination task only revealed marginal effects of training and the number of vowel contexts present in the discrimination task, it is possible that increasing variability in test (e.g., Fuhrmeister & Myers, 2017) and increasing trial-by-trial variability in the training task (e.g., Perrachione et al., 2011) may, in combination, have more detrimental effects on learning than one of those factors in isolation. To test this, we carried out an exploratory analysis, in which we collapsed the 2 × 2 group design to a single, four-level factor of group. We conducted a linear mixed effects model that included *d′* scores as the dependent variable and fixed effects of group (two-vowel test: interleaved training, one-vowel test: interleaved training, two-vowel test: blocked training, and one-vowel test: blocked training) and time (pretest, immediate posttest, and next-day posttest). Random effects included by-subject random intercepts. Factors in this model were dummy coded to test for simple effects, and the reference level was the two-vowel test: interleaved training group at the pretest. This allowed us to test whether the two sources of variability/difficulty in this study (training type and number of vowel contexts at test) would be more detrimental to learning or retention than just one at a time. This analysis indicated that the two-vowel test: interleaved training group's performance differed from the pretest at both the immediate posttest, β = 0.32, *SE* = 0.14, *t* = 2.30, *p* = .02, and the next-day posttest, β = 0.36, *SE* = 0.14, *t* = 2.58, *p* = .01. No significant differences

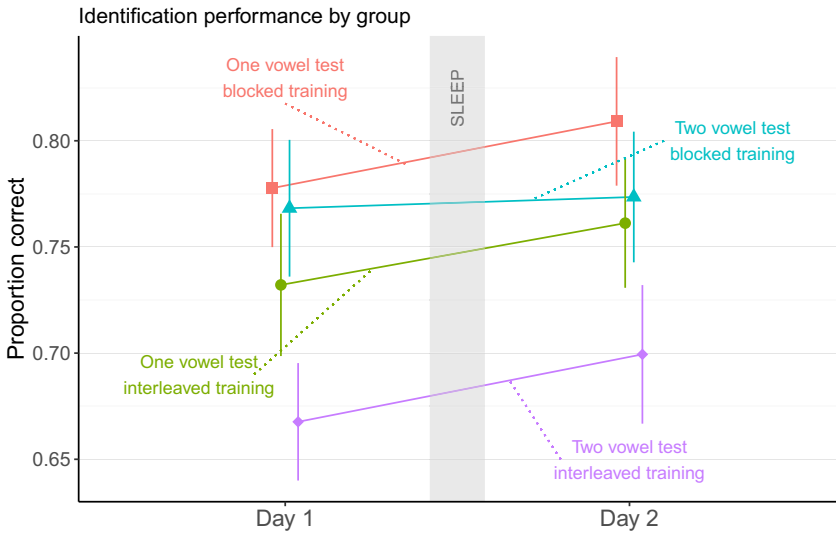

Fig. 3 Proportion correct responses on the posttraining identification assessment for each subgroup at each time point. The blocked training groups performed significantly better than the interleaved groups at each time point, and in general, performance was significantly better at Day 2. Error bars indicate standard error of the mean

were observed between the two-vowel test: interleaved training group and other groups at pretest. However, we observed an interaction with the immediate posttest time point and the one-vowel test: blocked training group, $\beta = 0.46$, $SE = 0.19$, $t = 2.36$, $p = .02$, as well as an interaction with the next-day posttest time point and the one-vowel test: blocked training group, $\beta = 0.43$, $SE = 0.19$, $t = 2.21$, $p = .03$ (see Table 6, Fig. 4). This suggests that the one-vowel test: blocked training group, the group with the lowest amount of (trial-by-trial) variability, was able to benefit more from training than the two-vowel test: interleaved training group, the group with the most (trial-by-trial) variability, and this advantage was retained from one day to the next.

Individual aptitude in phonetic learning and retention

To test the hypothesis that an individual's perceptual aptitude (as measured by pretraining discrimination scores) would predict learning or retention in training, identification tests, or discrimination tests and whether training (blocked vs. interleaved) or the number of vowels presented in test (one vs. two) had an effect on this relationship, we carried out a series of mixed effects models, described below.

Training To determine whether individual aptitude predicts performance during training, we carried out a mixed effects logistic regression model. In this model, the dependent

Table 4 Descriptive statistics for performance on discrimination tests by group and time point

Group	Time	N	dprime	sd
One vowel test: blocked training	Pretest	35	0.47	0.64
One vowel test: interleaved training	Pretest	33	0.33	0.59
Two vowel test: blocked training	Pretest	34	0.39	0.59
Two vowel test: interleaved training	Pretest	34	0.28	0.44
One vowel test: blocked training	Immediate posttest	35	1.25	1.13
One vowel test: interleaved training	Immediate posttest	33	0.86	0.88
Two vowel test: blocked training	Immediate posttest	34	0.84	1.14
Two vowel test: interleaved training	Immediate posttest	34	0.60	0.80
One vowel test: blocked training	Next-day posttest	35	1.26	1.21
One vowel test: interleaved training	Next-day posttest	33	0.94	1.12
Two vowel test: blocked training	Next-day posttest	34	1.01	1.13
Two vowel test: interleaved training	Next-day posttest	34	0.64	1.01

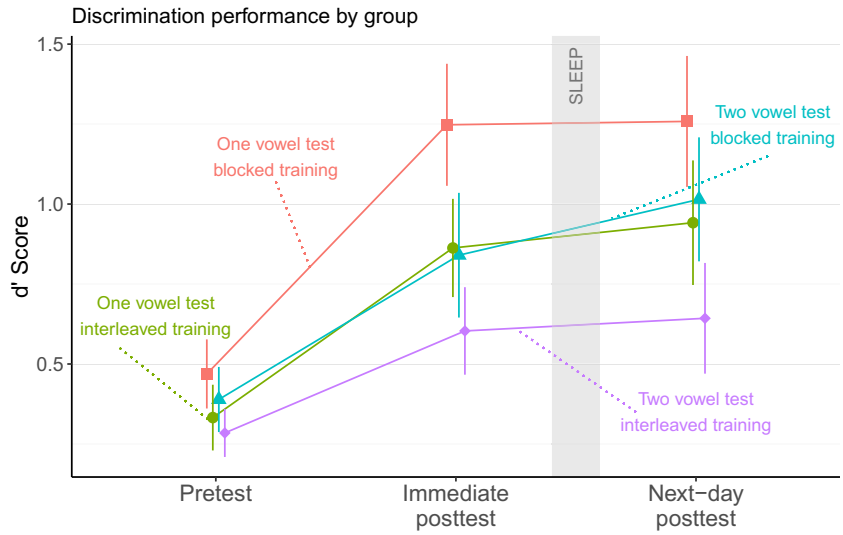

Fig. 4 Performance on the discrimination assessments (*d′* score) for each subgroup at each time point. The blocked training groups performed marginally better than the interleaved groups at each posttest. The one-vowel test blocked training group performed significantly better than the two-vowel test interleaved training group at each posttest. Error bars indicate standard error of the mean

variable was the log odds of selecting a correct response, and fixed effects included aptitude (pretraining discrimination score), training (blocked or interleaved), vowel context (/i/ or /u/), and number of vowels in the discrimination test (one or two). Categorical predictors were deviation coded (−.5, .5) as follows: training (blocked = −.5, interleaved = .5), vowel (/i/ = −.5, /u/ = .5), and test (one = −.5, two = .5). The random effects structure included by-subject random intercepts and slopes for vowel and by-item random intercepts and slopes for training, test, and their interaction. As in the previous analysis of performance on the training task, we observed a main effect of training, $\beta = -0.46$, $SE = 0.13$, $z = -3.67$, $p = .0002$, such that the blocked group was more accurate than the interleaved group in training. In addition, the analysis revealed a main effect of aptitude, $\beta = 0.65$, $SE = 0.09$, $z = 7.14$, $p < .0001$, suggesting that higher aptitude positively predicts performance during learning. We also found an interaction of vowel and aptitude, $\beta = 0.78$, $SE = 0.11$, $z = 6.85$, $p < .0001$, indicating that the relationship between aptitude and training performance was stronger for /u/ vowel trials than /i/ vowel trials. No other main effects or interactions were found (see Fig. 5).

Identification To test whether pretraining discrimination abilities contributed to learning or retention of the identification task, we conducted a mixed effects logistic regression model. As before, data were recoded for participants who appeared to have switched labels (see Identification Performance, above). The dependent variable in this model was the log odds of selecting a correct response, and fixed effects included aptitude (pretraining discrimination score), training (blocked or interleaved), number of vowel contexts at test (one or two), and time (Day 1 or Day 2). Categorical predictors were deviation coded (−.5, .5) as follows: training (blocked = −.5, interleaved = .5), test (one = −.5, two = .5), time (Day 1 = −.5, Day 2 = .5). The final model included by-subject random intercepts and slopes for time and random intercepts for item, and correlation parameters were set to zero. To get models to converge, we again used the optimizer "bobyqa" in the glmer control options and increased the number of iterations to 200,000.

This analysis revealed a main effect of aptitude, $\beta = 1.47$, $SE = 0.19$, $z = 7.565$, $p < .0001$, a main effect of training, $\beta = -0.56$, $SE = 0.23$, $z = -2.44$, $p = .015$ (with the blocked group showing better performance), and a main effect of time, $\beta = 0.16$, $SE = 0.07$, $z = 2.19$, $p = .03$, indicating that participants improved on the task from Day 1 to Day 2. In addition, we observed a significant interaction between aptitude and training, $\beta = 0.85$, $SE = 0.39$, $z = 2.20$, $p = .03$, and a nonsignificant trend towards an interaction of aptitude and time, $\beta = 0.29$, $SE = 0.15$, $z = 1.87$, $p = .06$. The pattern of the aptitude and training interaction was that the relationship between pretraining discrimination scores and performance on the identification posttest was stronger for those who received interleaved training. The trend towards an interaction between aptitude and time indicates that the relationship between pretraining discrimination and identification performance was slightly stronger on the second day, which provides weak evidence that higher pretraining aptitude predicts improvement after a period of off-line consolidation (see Fig. 6).

Discrimination To determine whether pretraining aptitude influenced learning or retention in the discrimination task, we carried out a linear mixed effects model. The dependent variable in this model was the *d′* posttest discrimination scores, and fixed effects included aptitude (pretraining discrimination

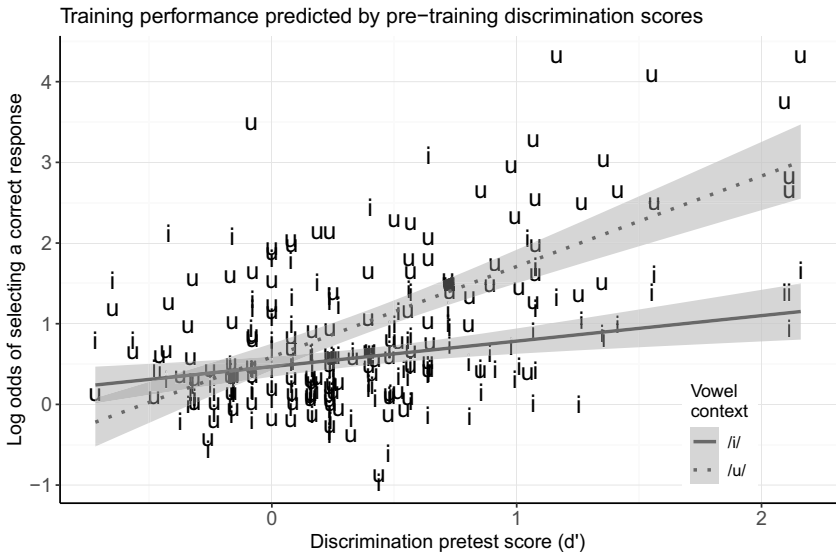

Fig. 5 Performance on the training task as predicted by aptitude (pretraining discrimination scores). Aptitude positively predicted accuracy on the training task, and this relationship was stronger for the /u/ vowel context than the /i/ vowel context. Blocked groups were also more accurate than interleaved groups; however, this is likely due to task differences (2AFC vs. 4AFC), so this is not shown here. For visualization purposes, percentage correct scores were averaged over participant and vowel context and converted to log odds so that each point above represents one participant's performance on the training task with trials of that vowel context. Shaded region indicates 95% confidence intervals

score), training (blocked or interleaved), number of vowel contexts at test (one or two), and time (immediate posttest or next-day posttest). Categorical predictors were deviation coded (−.5, .5) as follows: training (blocked = −.5, interleaved = .5), test (one = −.5, two = .5), time (immediate posttest = −.5, next-day posttest = .5). Random effects included by-subject random intercepts.

This analysis revealed a main effect of training, $\beta = -0.35$, $SE = 0.18$, $t = -1.99$, $p = .0489$, such that the blocked group showed better discrimination of the sounds than the interleaved group. In addition, we found a main effect of aptitude, $\beta = .99$, $SE = 0.14$, $t = 7.13$, $p < .0001$, suggesting that pretraining discrimination ability positively predicted posttraining discrimination. Finally, we observed a significant

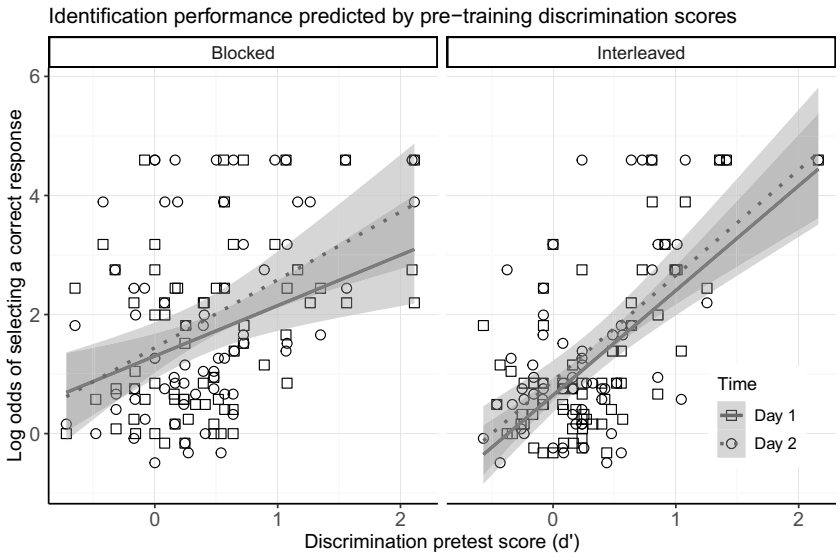

Fig. 6 Accuracy on the identification test as predicted by aptitude (pretraining discrimination scores) for blocked and interleaved groups at each time point. Aptitude positively predicted identification performance, but it interacted with training such that the relationship was stronger for the interleaved group than the blocked group. The blocked group was more accurate than the interleaved group, and Day 2 accuracy was significantly higher than Day 1 accuracy. For visualization purposes, percentage correct scores were averaged over participant and converted to log odds so that each point above represents one participant's performance on the training task with trials of that vowel context. For participants who were at 100% accuracy, these values were changed to .99 before converting to log odds to avoid infinite values. Shaded region indicates 95% confidence intervals

interaction between aptitude and time, β = 0.21, *SE* = 0.09, *t* = 2.46, *p* = .02, which indicates that the relationship between aptitude and the posttest score was stronger at the next-day posttest than the immediate posttest. This suggests that aptitude predicted improvement after a period of off-line consolidation (see Fig. 7).

Discussion

Blocked versus interleaved training

Variability is known to affect category learning. In the current study, we manipulated variability in both training and testing phases of nonnative phonetic learning. We first found that during training, participants who learned the novel speech sounds in a blocked design outperformed those who learned in an interleaved design, although the task design likely played a role in this (i.e., participants in the interleaved training design had to choose between four alternatives rather than two). To further examine the effects of seeing four versus two alternatives during training, we conducted a follow-up analysis, examining how errors were distributed during the interleaved (four-alternative) task. Interestingly, 89.5% of the time, an error consisted of picking a picture that matched the vowel of the correct picture, but with the wrong the consonant—that is, a participant, when hearing /d̪ug/ would erroneously choose the picture for /d̪ug/, but not /d̪ig/. This suggests that participants in general treated this task as a two-alternative task, easily narrowing down the possibilities to the two candidates that shared the same vowel. Although future manipulations are necessary to confirm this view, error data hints that the interleaved training penalty is not entirely attributable to the number of alternatives, but also to the demands of shifting trial-by-trial variability (e.g., Perrachione et al., 2011).

In addition (and unexpectedly), training performance was superior on trials in which the Hindi sounds were heard in the vowel context /u/ as compared with /i/. These two vowels differ acoustically in the second formant frequency (/i/ has a higher F2 value than /u/), which may have obscured the relevant acoustic cues to distinguish dental from retroflex sounds in the /i/ context. Although this was unexpected, pilot data from our lab suggest that this finding is robust even with stimuli recorded from other talkers and with different nonnative contrasts.

Contrary to predictions from the contextual interference effect (e.g., Battig, 1972; Magill & Hall, 1990; Shea & Morgan, 1979), we found that blocked training was related to better posttest performance (on identification, marginal for discrimination), and that these differences persisted after a delay. Importantly, although the training task differed between the blocked and interleaved conditions (2AFC vs. 4AFC), the identification test did not, with all participants only tested on their identification of the sounds presented in the /u/ vowel context. In addition, an exploratory analysis showed that the subgroup with the least amount of (trial-by-trial) variability (one-vowel test: blocked training) improved more on discrimination of the sounds after training (and maintained this advantage after a delay) than the subgroup exposed to the most (trial-by-trial) variability (two-vowel test: interleaved training). This suggests that each source of variability (an extra vowel context during test and interleaving practice trials) contributed to more difficult learning conditions than just one on its own; however, future research should test this question in a

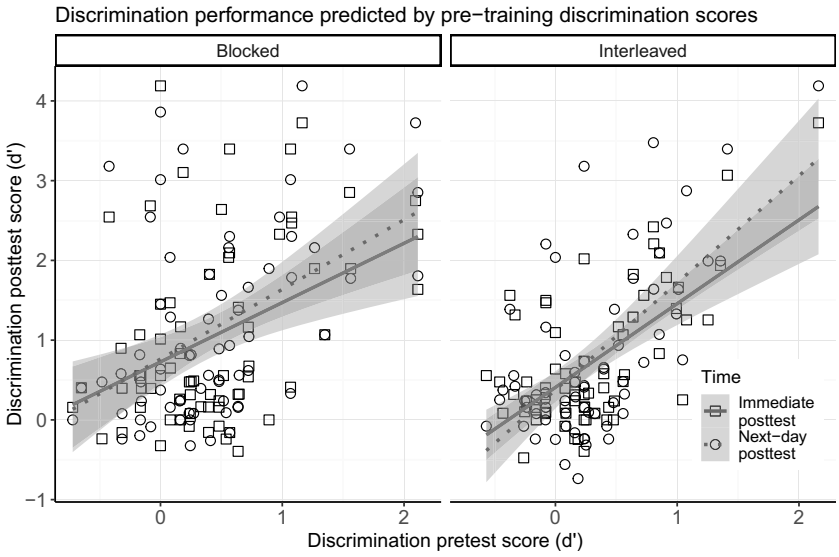

Fig. 7 Discrimination posttest scores as predicted by aptitude (pretraining discrimination score) shown for blocked and interleaved training groups. Blocked groups had higher posttest scores than interleaved groups, and aptitude positively predicted posttest scores, and this relationship was stronger for the next-day posttest than the immediate posttest. Shaded region indicates 95% confidence intervals

more hypothesis-driven way. Taken together, these results provide some evidence that blocking trials with varying phonological contexts during training of nonnative speech sounds leads to better learning and retention.

The contextual interference effect predicts that retention should be superior after more difficult learning conditions (i.e., interleaved practice; Magill & Hall, 1990; Shea & Morgan, 1979); however, we did not find that the interleaved groups showed superior retention compared with the blocked groups. The difficulty of learning the Hindi voiced dental and retroflex stop consonants for English-speaking adults may be responsible for the divergence from previous literature seen in our results. Learning phonetic contrasts that are perceptually similar to native-language speech sounds is particularly challenging for adults (e.g., Best & Tyler, 2007; Kuhl, 1994; Kuhl et al., 2008), and because the English alveolar /d/ sound encompasses allophonic variants of the Hindi dental /d̪/ and retroflex /ɖ/ (e.g., in "width" and "drip"; Polka, 1991), this phonetic contrast is one of the most difficult nonnative speech contrasts for native English speakers to learn (e.g., Best et al., 2001). Acquisition of this skill may have been too complex and therefore processing demands were too high for many learners to benefit from interleaved practice in our study. In this sense, interleaved presentation for already very difficult phonological contrasts may represent an "undesirable difficulty," resulting in poorer encoding of the contrast. Notably, however, our study design only included two vowel contexts, and previous studies testing the effects of high-variability training have used more variability in the stimulus set, including multiple talkers or multiple phonological contexts. It is nonetheless interesting that even with only one talker and two phonological contexts, we still see benefits of blocking training.

Our findings are inconsistent with several earlier studies that found advantages of high-variability training for phonetic learning (e.g., Bradlow et al., 1999; Bradlow et al., 1997; Lively et al., 1993; Lively et al., 1994; Logan et al., 1991); however, our study has some key differences that may explain the inconsistent findings. First, in the studies by Bradlow et al. (1999; Bradlow et al., 1997), Lively et al. (1993; Lively et al., 1994), and Logan et al. (1991), participants received far more training sessions than in our study. For example, in Bradlow et al. (1997), participants completed 45 training sessions over 3–4 weeks. Because we were primarily interested in the effects of variability in training and assessments of phonetic information after a period of off-line consolidation, participants completed only one training session and returned the following day for reassessment. Additionally, the participants in the aforementioned studies (native speakers of Japanese) were not naïve to the contrast: Many had been learning English for several years and probably had a great deal of exposure to the sounds before participating in the study and were likely very motivated to improve their perception of that contrast. Our participants, on the other hand, were completely unfamiliar with the Hindi contrast, which may have increased the difficulty and processing demands of the task to an even greater extent. This may explain why the participants who received blocked training had better posttraining identification scores: The lack of trial-by-trial variability may have reduced processing demands and allowed participants to learn the information more efficiently. It is possible that the processing demands were higher for the interleaved training groups because having four options on the screen was simply more difficult; however, it is also possible that the difficulty from the trial-by-trial acoustic variability did not allow learners to form categories. In either case, it seems that the 4AFC task that the interleaved groups completed was indeed more difficult. Therefore, our findings are still inconsistent with the contextual interference effect, in that more difficult learning conditions did not lead to superior retention after a delay. Future research could extend these findings with an additional interleaved training group that only sees the two alternatives relevant to the vowel context on each trial.

Phonological variability during unsupervised test

We also further explored a previous finding, that introducing variability in phonetic assessments hinders learning and retention (Fuhrmeister & Myers, 2017). As in Fuhrmeister and Myers (2017), half of participants in the present study heard two vowel contexts at test, and the other half heard only one. However, in contrast to Fuhrmeister and Myers, participants in the current study were trained on both of the vowel contexts they were exposed to during testing, which allowed us to consider the possibility that unsupervised variability (i.e., exposure to variability without feedback) was the source of this difficulty. We found no differences between the groups exposed to one or two vowel contexts in training or testing in the current study, which suggests that training on the second vowel context can indeed mitigate these detrimental effects on learning seen in our previous study. Thus, variability introduced in phonetic training may need to occur in the context of supervised learning (i.e., with feedback); otherwise, it obfuscates the acoustic cues necessary to distinguish the speech sounds. This finding generates the hypothesis that supervised, but not unsupervised variability in phonological contexts can benefit nonnative phonetic learning. It is also worth noting that the two-vowel groups were exposed to twice as many tokens in the discrimination tests than the one vowel groups (128 tokens vs. 64 tokens). As in Fuhrmeister and Myers, it is surprising that this additional exposure to the Hindi sounds did not result in stronger learning or retention for this group, even when given explicit feedback on the sounds presented in both vowel contexts during the training task. This finding lends further support to the idea that unsupervised variability may not be beneficial, at least in the early stages of learning.

Individual aptitude

Unsurprisingly, pretraining discrimination ability significantly predicted performance on the training task and the identification and discrimination posttests, showing that participants who were better able to detect differences between the dental and retroflex stimuli before training went on to learn the contrast more robustly.

This relationship between aptitude and performance on the identification test was even stronger for the interleaved training group than the blocked training group, suggesting that the higher-aptitude participants were best able to take advantage of the challenging interleaved training condition. This is consistent with findings from Perrachione et al. (2011) and Sadakata and McQueen (2014) and provides further evidence that variability in training may not be beneficial to every learner. This finding is also in line with the notion of desirable difficulties (Bjork, 1994) and suggests that learning is most effective when an individual is provided an appropriate level of difficulty for his or her skill level. Notably, we did not see this interaction in the discrimination analyses. It is possible that because the identification test was very similar to training (except it did not include feedback and only included the /u/–vowel context), high-aptitude learners were better able to capitalize on the interleaved training when no generalization to a different task was necessary.

One motivator of the current study was the possibility that individuals may differ not only in their ability to learn new sounds but also the degree to which off-line consolidation processes solidify (or enhance) learned categories in memory. We found that the relationship between aptitude and phonetic discrimination was stronger after a period of off-line consolidation compared with before. In other words, aptitude predicted overnight improvement of discrimination of the sounds. This pattern of results was limited to the discrimination test (although this interaction was approaching significance in the analyses of the identification test), which could be because discrimination was not explicitly trained and as such, discrimination can be thought of as a test of generalization to a new task (see Earle & Myers, 2014).

This finding is consistent with the literature on sleep and memory consolidation (see Earle & Myers, 2014, for review). Although we did not specifically test whether sleep played a role per se (i.e., we did not have any groups of participants who had an equivalent waking interval between training and delayed posttests), our study design included an overnight interval for all participants. Furthermore, previous studies have explicitly tested the role of sleep in learning nonnative speech sounds and have found that learners generally improve after a period of sleep (Earle, Landi, & Myers, 2017; Earle & Myers, 2015b) and that learners are able to generalize their knowledge of newly learned phonetic contrasts to an untrained talker after sleeping (Earle & Myers, 2015a). The complementary learning systems theory predicts that learned information is abstracted away from episodic details during sleep; therefore, learners should show better generalization after sleeping (McClelland, McNaughton, & O'Reilly, 1995). Some work from the sleep and memory consolidation literature additionally suggests that information that is more strongly encoded is preferentially consolidated (Ebbinghaus, 1885; Hauptmann et al., 2005; Shibata et al., 2017; Tucker & Fishbein, 2008). The fact that aptitude predicted overnight improvement on discrimination in the current study is consistent with this idea: Higher-aptitude participants likely encoded the information more strongly and therefore showed more overnight improvement on a task of generalization than participants with lower perceptual aptitudes. Perhaps this relationship in the identification test did not reach significance because there was no generalization needed to perform that task.

One unanticipated finding was differential effects according to the vowel context itself. In the training task, we observed an interaction between aptitude and vowel context, showing that the relationship between aptitude and accuracy on the training task was stronger for /u/ trials as compared with /i/ trials. Accuracy was lower on /i/–vowel trials than /u/–vowel trials, regardless of whether participants received blocked or interleaved training or were exposed to one or two vowel contexts in test. Some previous studies suggest that aptitude should benefit learners in more difficult learning situations. For example, the notion of "desirable difficulties" (e.g., Bjork, 1994) posits that learners learn best when they are faced with an optimal level of difficulty (i.e., not too easy but not so difficult that they are unable to learn). Thus, high-aptitude learners should benefit from more difficult training conditions. In addition, Sadakata and McQueen (2014) found that learners of a Mandarin tonal contrast with higher pretraining aptitude fared better with high-variability training than low-aptitude learners, but this relationship did not hold for an (arguably) easier phonological contrast, specifically, a geminate contrast. This provides further support for the idea that aptitude, or pretraining ability for a task, is most beneficial when the learning task is more difficult. This is not what we found with the interaction of aptitude and vowel context on the training task in the current study; however, as discussed earlier, the acoustic properties of the /i/ vowel may have obfuscated the critical acoustic cues necessary for distinguishing the dental and retroflex sounds and may have been too challenging even for high aptitude learners.

Our results add to the literature suggesting that phonetic training with variability comes at a cost, at least for some learners. However, variability has been shown to promote generalization or transfer of information to novel contexts in phonetic learning (i.e., sounds produced by novel talkers or occurring in different phonological contexts, Bradlow et al., 1999; Bradlow et al., 1997; Lively et al., 1993; Lively et al.,

1994; Logan et al., 1991), and generalization is indeed the ultimate goal of many types of learning. For example, a language learner in the real world must be able to quickly and efficiently map acoustic input onto speech categories in order to recognize words and derive meaning from an utterance. In the present study, we did not test for generalization to a novel talker or phonological context, and it is possible that we would see superior generalization in the group that received interleaved training. However, Perrachione et al. (2011) found no differences in generalization ability for learners who received high-variability training in blocked or interleaved conditions, so we are doubtful that we would have found better generalization for the interleaved group. In addition, because all groups received equal exposure to each vowel context in training, better generalization would need to stem from the increased difficulty in training that was a result of interleaving the different vowel contexts. Although such a finding is well predicted by the contextual interference effect, the contextual interference effect would similarly predict superior retention after more difficult learning conditions, but our interleaved groups did not retain the new phonetic information any better than the blocked groups (Magill & Hall, 1990; Shea & Morgan, 1979). Therefore, it is unclear whether the increased difficulty in training caused by interleaving different types of trials is beneficial for phonetic learning. We might similarly predict that any advantages for generalization due to interleaved training would be limited to learners with better perceptual abilities. Phonetic learning, especially learning of particularly difficult speech sound contrasts, may be too difficult for many learners to take advantage of random or interleaved practice. We have argued that learning needs to be strong enough in order to trigger off-line consolidation processes, especially those that occur during sleep, so it is crucial for learners to reach a critical level of stability in their initial learning to more effectively consolidate new phonetic information. Consistent with findings from Perrachione et al. (2011) and Sadakata and McQueen (2014), our results suggest that one way to achieve this is to expose learners with poorer perceptual abilities to variability in a blocked manner. Ultimately, striking a delicate balance between variability and an optimal level of training difficulty for a given individual is key to promoting more efficient learning and retention of new phonetic categories.

Acknowledgements This material is based upon work supported in part by the National Science Foundation under Grant DGE-1747486, NSF IGERT DGE-1144399 to the University of Connecticut, and NSF BCS 1554510 to E.B.M. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The data and analysis scripts are available at https://osf.io/ujm4f/, and the experiment was not preregistered.

References

Antoniou, M., & Wong, P. C. (2015). Poor phonetic perceivers are affected by cognitive load when resolving talker variability. *The Journal of the Acoustical Society of America*, *138*(2), 571–574.

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105–1138.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*(6), 1178–1199.

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

Battig, W. F. (1972). *Interference during learning as a source of facilitation in subsequent retention and transfer.* Paper presented at 'single presentation series' of Division C (Instruction and Learning) at the annual meeting of the American Educational Research Association, Chicago, IL.

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

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), Language experience in second language speech learning: In honor of James Emil Flege (pp. 13–34). Philadelphia, PA: John Benjamins. https://doi.org/10.1075/lllt.17.07bes

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), Metacognition: Knowing about knowing (pp. 185–205). Cambridge, MA: MIT Press.

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer (Version 6.0. 37) [Computer software]. Retrieved from http://www.fon.hum.uva.nl/praat/

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. I. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977–985.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to nonnative speech. *Cognition, 106*(2), 707–729.

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.

Brown, H., & Gaskell, M. G. (2014). The time-course of talker-specificity and lexical competition effects during word learning. *Language, Cognition and Neuroscience*, *29*(9), 1163–1179.

Chandrasekaran, B., Koslov, S. R., & Maddox, W. T. (2014a). Toward a dual-learning systems model of speech category learning. *Frontiers in Psychology*, *5*, 825.

Chandrasekaran, B., Yi, H. G., & Maddox, W. T. (2014b). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, *21*(2), 488–495.

Drosopoulos, S., Schulze, C., Fischer, S., & Born, J. (2007). Sleep's function in the spontaneous recovery and consolidation of memories. *Journal of Experimental Psychology: General*, *136*(2), 169.

Earle, F. S., Landi, N., & Myers, E. B. (2017). Sleep duration predicts behavioral and neural differences in adult speech sound learning. *Neuroscience Letters*, *636*, 77–82.

Earle, F. S., & Myers, E. B. (2014). Building phonetic categories: an argument for the role of sleep. *Frontiers in Psychology*, *5*, 1192.

Earle, F. S., & Myers, E. B. (2015a). Overnight consolidation promotes generalization across talkers in the identification of nonnative

speech sounds. *The Journal of the Acoustical Society of America*, *137*(1), EL91–EL97.

Earle, F. S., & Myers, E. B. (2015b). Sleep and native language interference affect nonnative speech sound learning. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1680.

Ebbinghaus, H. (1885). Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie [About memory: Investigations on experimental psychology]. Leipzig, Germany: Verlag von Duncker und Humblot.

Ell, S. W., & Ashby, F. G. (2012). The impact of category separation on unsupervised categorization. *Attention, Perception, & Psychophysics*, *74*(2), 466–475.

Fuhrmeister, P., & Myers, E. B. (2017). Nonnative phonetic learning is destabilized by exposure to phonological variability before and after training. *The Journal of the Acoustical Society of America*, *142*(5), EL448-EL454.

Galle, M. E., Apfelbaum, K. S., & McMurray, B. (2015). The role of single talker acoustic variation in early word learning. *Language Learning and Development*, *11*(1), 66–79.

Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: Reallocation of neural substrates. *NeuroImage*, *21*(2), 494–506.

Goudbeek, M., Cutler, A., & Smits, R. (2008). Supervised and unsupervised learning of multidimensionally varying nonnative speech categories. *Speech Communication*, *50*(2), 109–125.

Hall, K. G., Domingues, D. A., & Cavazos, R. (1994). Contextual interference effects with skilled baseball players. *Perceptual and Motor Skills*, *78*(3), 835–841.

Hauptmann, B., Reinhart, E., Brandt, S. A., & Karni, A. (2005). The predictive value of the leveling off of within session performance for procedural memory consolidation. *Cognitive Brain Research*, *24*(2), 181–189.

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

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393–1409.

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

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

Li, Y., & Wright, D. L. (2000). An assessment of the attention demands during random-and blocked-practice schedules. *The Quarterly Journal of Experimental Psychology Section A*, *53*(2), 591–606.

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

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3), 1242–1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y. I., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *The Journal of the acoustical society of America*, *96*(4), 2076–2087.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, *89*(2), 874–886.

Macmillan, N. A., & Creelman, C. D. (2005). Detection theory: A user's guide (2nd ed.). Mahwah, NJ: Erlbaum.

Magill, R. A., & Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human Movement Science*, *9*(3/5), 241–289.

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

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience*, *2*(2), 89–108.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378.

Myers, E. B., & Swan, K. (2012). Effects of category learning on neural sensitivity to nonnative phonetic categories. *Journal of Cognitive Neuroscience*, *24*(8), 1695–1708.

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

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

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

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology*, *5*, 1318.

Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr. (1998). Contextual interference effects in foreign language vocabulary acquisition and retention. In A. F. Healy & L. E. Bourne, Jr. (Eds.), Foreign language learning: Psycholinguistic studies on training and retention (pp. 77–90). Mahwah, NJ: Erlbaum.

Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*(2), 419–440.

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(2), 179.

Shibata, K., Sasaki, Y., Bang, J. W., Walsh, E. G., Machizawa, M. G., Tamaki, M., Chang, L., & Watanabe, T. (2017). Overlearning hyperstabilizes a skill by rapidly making neurochemical processing inhibitory-dominant. *Nature Neuroscience, 20,* 470–475.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2019). afex: Analysis of factorial experiments (R Package Version 0.23-0) [Computer software]. Retrieved from https://CRAN.R-project.org/package=afex

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

Tucker, M. A., & Fishbein, W. (2008). Enhancement of declarative memory performance following a daytime nap is contingent on strength of initial task acquisition. *Sleep*, *31*(2), 197–203.

Vlahou, E. L., Protopapas, A., & Seitz, A. R. (2012). Implicit training of nonnative speech stimuli. *Journal of Experimental Psychology: General*, *141*(2), 363.

Wade, T., & Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *The Journal of the Acoustical Society of America*, *118*(4), 2618–2633.

Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review*, *9*(2), 185–211.

Yi, H. G., Maddox, W. T., Mumford, J. A., & Chandrasekaran, B. (2014). The role of corticostriatal systems in speech category learning. *Cerebral Cortex*, *26*(4), 1409–1420.