Boosting lexical support does not enhance lexically guided perceptual learning

Sahil Luthra[1,2], James S. Magnuson[1,2], & Emily B. Myers[1,2,3]

[1] Department of Psychological Sciences, University of Connecticut

[2] The Connecticut Institute for the Brain and Cognitive Sciences

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

Contact

sahil.luthra@uconn.edu (S. Luthra, corresponding author)

james.magnsuon@uconn.edu (J. S. Magnuson)

emily.myers@uconn.edu (E. B. Myers)

Author Note

This research was supported by NIH R01 DC013064 (E.B.M., PI), by NSF IGERT grant DGE-1144399 (J.S.M., PI), and by NSF Research Traineeship grant (NRT) IGE1747486 (J.S.M., PI), and NSF 1754284 (J.S.M., PI). SL was supported by an NSF Graduate Research Fellowship. The authors thank the members of the Cognitive Neuroscience of Language Lab, the Language and Brain Lab, and the Spoken Language Processing lab for their feedback throughout the project. We thank Rachael Steiner for her assistance in programming the experiment as well as for feedback on a previous version of this manuscript. We also thank Gerry T. M. Altmann and Rachel M. Theodore for their feedback on a previous version of this manuscript. All analysis scripts are available at https://osf.io/eqwja/.

Abstract

A challenge for listeners is to learn the appropriate mapping between acoustics and phonetic categories for an individual talker. Lexically guided perceptual learning (LGPL) studies have shown that listeners can leverage lexical knowledge to guide this process. For instance, listeners learn to interpret ambiguous /s/-/ʃ/ blends as /s/ if they have previously encountered them in /s/-biased contexts like *epi?ode*. Here, we examined whether the degree of preceding lexical support might modulate the extent of perceptual learning. In Experiment 1, we first demonstrated that perceptual learning could be obtained in a modified LGPL paradigm where listeners were first biased to interpret ambiguous tokens as one phoneme (e.g., /s/) and then later as another (e.g., /ʃ/). In subsequent experiments, we tested whether the extent of learning differed depending on whether targets encountered predictive contexts or neutral contexts prior to the auditory target (e.g., *epi?ode*). Experiment 2 used auditory sentence contexts (e.g., *I love "The Walking Dead" and eagerly await every new…*), while Experiment 3 used written sentence contexts. In Experiment 4, participants did not receive sentence contexts but rather saw the written form of the target word (*episode*) or filler text (########) prior to hearing the critical auditory token. While we consistently observed effects of context on in-the-moment processing of critical words, the size of the learning effect was not modulated by the type of context. We hypothesize that boosting lexical support through preceding context may not strongly influence perceptual learning when ambiguous speech sounds can be identified solely from lexical information.

Keywords: Sentence context, lexical support, phonetic recalibration, perceptual learning, speech

A core principle in psychology is that perception is guided by past experience. In his classic paper, Goldstone (1998) described *perceptual learning* as a process by which some stimulus dimensions in the environment are highlighted and others are deemphasized, inducing "relatively long-lasting changes" in perception. Consider, for instance, the often-formidable challenge of trying to understand the speech of an unfamiliar talker, perhaps a person with a novel accent or a speech motor impairment. As listeners gain familiarity with that talker, they can make perceptual adjustments that facilitate comprehension in future encounters with that talker relative to an unfamiliar talker (e.g., Nygaard, Sommers, & Pisoni, 1994). Psycholinguists often refer to this form of perceptual learning for speech as *phonetic recalibration*, as it involves reconfiguring how the incoming acoustic signal maps onto known speech sounds (i.e., phonetic categories corresponding to spoken consonants and vowels) for a given talker.

Listeners are able to avail themselves of a variety of contextual cues in order to resolve ambiguity during spoken language comprehension, and such context can also be used to guide phonetic recalibration. For instance, a listener's interpretation of an acoustically ambiguous token is influenced by their lexical knowledge of which strings of sounds constitute real words (Ganong, 1980) as well as by the visible movements of a talker's articulators, such as the lips and jaw (McGurk & MacDonald, 1976). These contextual cues not only affect perception in the moment but can also be used by listeners to form inferences about how a talker tends to realize speech categories (Bertelson, Vroomen, & De Gelder, 2003; Norris, McQueen, & Cutler, 2003). In a seminal study, Norris et al. introduced a paradigm for *lexically guided perceptual learning* (LGPL; a specific case of phonetic recalibration in which lexical knowledge guides perceptual learning). Their findings have been replicated and extended in a number of similar experiments. In such studies, listeners typically hear a talker who produces a speech sound that is ambiguous between

two speech sounds, such as 's'-/s/ and 'sh'-/ʃ/; this ambiguous sound is denoted here as /?/. During an initial exposure phase, /?/ is only encountered in disambiguating lexical contexts; some participants only hear /?/ in contexts where lexical information guides them to interpret it as /s/ (e.g., *epi?ode*), whereas other listeners only hear /?/ in /ʃ/-biased contexts (e.g., *flouri?ing*). All participants also hear clear productions of the contrastive phoneme (e.g., participants who hear *epi?ode* also hear *flourishing*). Following exposure, participants are given a phonetic categorization task using sounds on a continuum from /s/ to /ʃ/; critically, no lexically disambiguating information is provided at test. Listeners who were exposed to /?/ in /s/-biased contexts categorize more of the continuum as /s/, whereas listeners who heard the /?/ in /ʃ/ contexts categorize more of the continuum as /ʃ/. That is, listeners adjust how they map that talker's acoustics onto phonetic categories on the basis of lexical context in the exposure phase. In their study, Norris et al. found that this perceptual learning did not occur when both possible completions of the exposure items resulted in nonwords, as there was no contextual information available to bias a subject's interpretation of the ambiguous segment. However, other studies have shown that perceptual learning can in principle occur after auditory exposure to nonwords, so long as there is some cue to the identity of the ambiguous segment, whether this cue comes from phonotactic rules (Cutler, McQueen, Butterfield, & Norris, 2008), distributional properties of the acoustics (Chládková, Podlipskỳ, & Chionidou, 2017), simultaneous presentation of written text (Keetels, Schakel, Bonte, & Vroomen, 2016), or coincident facial movements (Bertelson et al., 2003).

Notably, LGPL need not entail a general change in how all incoming acoustic information is mapped onto phonetic categories; rather, it generally involves a change in a listener's understanding how a *particular* talker produces their speech. As such, perceptual adjustments do

not tend to generalize to a different talker, at least for fricatives and vowels (Eisner & McQueen, 2005; Kraljic & Samuel, 2006, 2007); this is likely because there tends to be a considerable amount of variability in how talkers produce these particular sounds relative to the degree of variability in how they realize plosive consonants (Kleinschmidt, 2019; van der Zande, Jesse, & Cutler, 2014). Furthermore, the extent of perceptual learning appears to depend on whether the acoustic variation is characteristic of the talker; lexically guided perceptual learning of ambiguous fricatives has been shown not to occur if listeners initially hear the talker producing those sounds clearly (Kraljic, Brennan, & Samuel, 2008; Kraljic & Samuel, 2011; Kraljic, Samuel, & Brennan, 2008), though phonetic recalibration studies using ambiguous plosive consonants have successfully shifted a subject's category boundary in one direction and then in another (Bonte, Correia, Keetels, Vroomen, & Formisano, 2017; Kilian-Hütten, Valente, Vroomen, & Formisano, 2011; van Linden & Vroomen, 2007). Similarly, LGPL is attenuated when a talker has a pen in their mouth as they produce the ambiguous segment (Kraljic & Samuel, 2011; Kraljic, Samuel, & Brennan, 2008) because in such a situation, there is uncertainty about whether the atypical acoustic information should be attributed to the talker or to the pen (Liu & Jaeger, 2018). In this way, perceptual learning reflects a listener's inferences about the likely cause of an ambiguous acoustic signal.

In order for listeners to learn how to map an ambiguous acoustic signal onto a perceptual category, listeners must first encode the ambiguity. If the signal is not ambiguous, as at the clear endpoints of a phonetic continuum, learning is not observed (e.g., Kraljic & Samuel, 2005). Furthermore, if the acoustic input is encountered in difficult processing conditions, as when the critical items are presented in the context of speech-shaped noise, learning is attenuated (Zhang & Samuel, 2014). Similarly, work by Samuel (2016) suggests that the size of perceptual learning effects is reduced when attentional resources are directed away from the critical auditory

information. In a dichotic listening experiment, Samuel presented listeners with two voices – a female talker whose speech contained an ambiguous segment that was disambiguated by lexical context, and a male talker who interrupted the female talker close to the critical speech segment. When listeners were asked to attend to the female talker, they showed robust perceptual learning, but when listeners were asked to attend to the male voice, they did not show perceptual learning for the female talker. Samuel concluded that some degree of attention to the critical signal is therefore needed in order for perceptual learning to occur. A follow-up repetition priming experiment indicated that the lack of attention to the critical segment also impeded lexical access to the critical word, suggesting that attention to the signal may modulate whether lexical access occurs and that lexical access may in turn drive perceptual learning. Overall, then, findings suggest that the degree to which listeners can encode and attend to ambiguity in the acoustic signal modulates whether phonetic recalibration occurs.

At the same time, overt attention to acoustic-phonetic ambiguities has been shown to attenuate perceptual learning. In two LGPL studies, McAuliffe and Babel (2015, 2016) told one group of subjects that they would hear a talker who produced an ambiguous /s/, while a second group was not told explicitly about the ambiguity. The authors found that perceptual learning was attenuated when subjects' attention was explicitly called toward the ambiguity. They argued that under typical conditions, listeners process the acoustic signal in a comprehension-oriented way that facilitates perceptual learning; if listeners attend specifically to sub-lexical details, however, learning does not occur. Data from an individual differences study by Scharenborg, Weber and Janse (2015) also support the proposal that overtly attending to phonetic information may attenuate learning. In particular, Scharenborg et al. measured the attention-switching abilities of a sample of older adults by use of the Trail Making Test, in which individuals have to constantly switch

attention between written numbers and letters (Bowie & Harvey, 2006). They hypothesized that older adults with worse attentional switching abilities would be impaired in switching between top-down cues and bottom-up acoustic information and might therefore rely more heavily on the lexicon for top-down support during speech perception. Because these individuals were attending less to the acoustic ambiguities, they might be expected to show more perceptual learning. This prediction was borne out in their data and also observed in a separate sample of older adults by Colby, Clayards, and Baum (2018). Thus, data from several studies suggest that overt attention to sub-lexical details might in fact be deleterious for phonetic recalibration.

While perceptual learning can in principle be guided by many sources of context, including lipread information (van Linden & Vroomen, 2007) and written text (Bonte et al., 2017; Keetels et al., 2016; Mitterer & McQueen, 2009), it appears that the timing of this context plays a critical role in determining whether perceptual learning will occur. Theoretical accounts of perceptual learning posit that in order for learning to occur, contextual information must be available at the same time as (or perhaps very shortly after) the critical segment is processed (Davis & Johnsrude, 2007; Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008; Jesse & McQueen, 2011). In LGPL specifically, recalibration has been robustly observed when phonetic ambiguities are in word-medial (Kraljic & Samuel, 2005) or word-final positions (Jesse & McQueen, 2011; McAuliffe & Babel, 2015) but not when ambiguous phonemes are in word-initial positions (Jesse & McQueen, 2011; McAuliffe & Babel, 2015). The dichotic listening experiment conducted by Samuel (2016) found that a competing auditory stream could disrupt recalibration if it began no more than a second after the onset of the critical auditory information, suggesting that phonetic representations might be malleable for approximately one second. Thus, learning is observed only when contextual information is

available at roughly the same time as an ambiguous segment is encountered, but not if the disambiguating information comes some amount of time (possibly more than one second) after the ambiguous phoneme.

Consistent with this view, Jesse (2020) demonstrated that phonetic recalibration can occur when listeners encounter ambiguous fricatives in word-initial position if a preceding sentential context disambiguates the phoneme's identity. In her study, listeners encountered a fricative that was ambiguous between /s/ and /f/ in contexts where lexical information could not resolve phoneme identity (e.g., [*s/f*]*ame*, where both *same* and *fame* are real words). However, during their initial exposure phase, they passively listened to sentence contexts (e.g. *That she had met Lady Gaga at the airport would be her only claim to…*) that guided interpretation of the sentence-final word (here, *fame*), thus allowing recalibration to occur. Notably, no perceptual learning was observed when a neutral sentence context was used (e.g., *Click on the word…*), as such context could not guide interpretation of the final word ([*s/f*]*ame*). While these findings suggest that early contextual support is needed to guide perceptual learning in cases where the phonetic ambiguity results in lexical ambiguity (*same* vs. *fame*), they leave open questions of how the specific *degree* of contextual support modulates the extent of recalibration.

In the current study, we investigated how the strength of a listener's prior expectations might influence the extent of phonetic recalibration obtained from disambiguating lexical context. Specifically, we exposed participants to words with ambiguous medial segments (e.g., *epi?ode*), manipulating how strongly preceding contexts predicted the target words. Participants received either highly predictive contexts (e.g., *I love "The Walking Dead" and eagerly await every new…*) or contexts that were neutral with respect to the final word (*My ballpoint pen ran out of ink when I was halfway through writing the word…*). Note that in the current study, unlike in the work of

Jesse (2020), the full lexical context is sufficient to unambiguously identify the intended phoneme (as *episode* is a word and *epishode* is not, whereas both *same* and *fame* are real words). The critical variable in the current study is the extent to which the word (*episode*) is predicted by a preceding context. That is, does manipulating how strongly a preceding context predicts a particular lexical item affect how that lexical item guides perceptual learning?

It is unclear from the extant literature exactly how preceding context should influence LGPL. One possibility comes from *ideal observer* accounts (e.g., Kleinschmidt & Jaeger, 2015), which are based on principles of Bayesian inference and posit that listeners should be able to make use of every available source of information to constrain the phonetic categorization of a segment. Because a predictive context boosts the prior probability that a particular phoneme will be encountered, listeners have a stronger cue for the identity of an ambiguous phoneme that is preceded by a predictive context relative to a more neutral one. Thus, ideal observer accounts predict that LGPL should be enhanced when critical items follow predictive contexts relative to when they follow neutral contexts.

An alternative hypothesis comes from accounts that describe perceptual learning as a result of changes in *attentional weighting* (e.g., Goldstone, 1998), whereby stimulus dimensions that are more informative become more heavily weighted than stimulus dimensions that are less informative. When applied to phonetic recalibration, an attentional weighting account would argue that a predictive context permits listeners to resolve the identity of an ambiguous phoneme without attending strongly to the bottom-up signal. As such, listeners who hear predictive contexts should increase the attentional gain on lexical information (as described, for instance, by Pitt & Szostak, 2012) at the expense of the attention directed toward the acoustic signal. An attentional weighting account thus predicts that the degree of perceptual learning should be *reduced* when critical items

follow a predictive context relative to when they follow a neutral one. In a similar vein, studies have demonstrated that listeners are less likely to encode phonetic details when they occur in highly predicted words. A verbal shadowing experiment conducted by Marslen-Wilson and Welsh (1978) found that listeners were more likely to fluently restore mispronunciations if they occurred in neutral contexts (*It was his misfortune that they were stationary*) compared to in predictive ones (*Still, he wanted to smoke a cigarette*).

More recently, in a study by Manker (2019), listeners heard a sentence context that was either predictive (*The vampires are sleeping in...*) or not predictive (*The first thing Mary saw was the...*) of a target word (*coffins*). Following this sentence, listeners heard the target word (*coffins*) a second time and completed an AX discrimination task to indicate whether the second token of the target word was identical to the first. Manker found that subjects were significantly better at discriminating between productions of the target word when they followed a neutral context than when they followed a predictive one, therefore suggesting that context may modulate how much listeners attend to phonetic details. When applied to the current study, these data indicate that listeners who hear a critical item in a predictive context may not encode its phonetic details very well. As described above, previous LGPL studies have suggested that attention to the acoustic-phonetic details of the speech signal can have consequences for phonetic recalibration, with a reduction in recalibration observed when listeners' attention is directed to a competing speech stream (Samuel, 2016) and a reduction in recalibration also observed when attention is overtly directed toward an ambiguity (McAuliffe & Babel, 2015, 2016; Scharenborg et al., 2015). As such, strong predictions about the identity of an upcoming word may influence the degree to which listeners attend to the critical speech segment, potentially attenuating perceptual learning.

Here, we present the results of four experiments assessing whether the predictive power of preceding context influences the degree of LGPL. In Experiment 1, we measured the extent of perceptual learning when critical words were presented in isolation (i.e., without any preceding context). In Experiment 2, participants heard auditory sentence contexts prior to the critical words, and we assessed whether learning was larger for participants who received sentence contexts that were predictive of the critical word or for participants who received contexts that were neutral with respect to their predictive power. In Experiment 3, we used written contexts instead of auditory ones, allowing us to eliminate potential interference from auditory exposure to other fricatives (e.g., /z/, /ʒ/) in our sentence contexts. Finally, in Experiment 4, one group of participants saw the written form of the auditory target prior to hearing it, while another group saw filler text, allowing us to make contexts maximally or minimally predictive of the upcoming target stimulus.

Experiment 1

Experiment 1 was designed to provide a baseline estimate of the extent of perceptual learning when no preceding context was provided at exposure; in doing so, we also would verify that phonetic recalibration would be obtained with slight modifications to the standard LGPL paradigm. In particular, we anticipated that it might seem unnatural if nonwords were presented after sentence contexts, and so we opted to present only words during exposure. As such, we used a semantic categorization task (judging whether the word was a concrete noun) during exposure instead of the standard lexical decision task (e.g., Kraljic & Samuel, 2005; Norris et al., 2003); learning was assessed using a phonetic categorization task (categorizing stimuli along a "sign"–"shine" continuum). While the semantic categorization task used in our exposure phase has not, to our knowledge, been used in previous LGPL studies, recalibration has been observed following a

broad range of exposure tasks, such as counting the number of words heard during exposure (McQueen, Norris, & Cutler, 2006), old/new judgments (Leach & Samuel, 2007), same/different judgments (Clarke-Davidson, Luce, & Sawusch, 2008), loudness judgments (Drouin & Theodore, 2018), syntactic judgments (Drouin & Theodore, 2018), and even passive exposure (Eisner & McQueen, 2006; Maye, Aslin, & Tanenhaus, 2008; White & Aslin, 2011). This suggests that the phonetic recalibration can be induced by a myriad of exposure tasks, so long as the task encourages the listener to resolve the ambiguous sounds to the intended phonetic category (Kleinschmidt & Jaeger, 2015).

Additionally, a standard LGPL paradigm exposes a given participant to the ambiguous token in only one biasing condition; for instance, a subject might either hear /?/ in /s/-biased contexts or in /ʃ/-biased contexts, but not both (i.e., Bias is treated as a between-subjects factor). Following van Linden and Vroomen (2007), we instead manipulated Bias within subjects, first providing subjects with one set of biasing contexts (e.g., /s/-biased contexts), assessing their learning with a phonetic categorization task, and then repeating the procedure with the opposite set of biasing contexts (e.g., /ʃ/-biased contexts); this procedure is schematized in Figure 1.

If participants recalibrate on the basis of their previous exposure (as expected from previous LGPL studies), we should see an effect of the most recent Bias condition (/s/ or /ʃ/) on participants' responses during the phonetic categorization task. We also expect an effect of which Step along the continuum participants are hearing, with participants making more /ʃ/ responses when presented with more /ʃ/-like tokens. Since perceptual learning should not be observed for clear continuum endpoints but should be observed for ambiguous stimuli, we may see a Bias × Step interaction, though one would not be required for evidence of phonetic recalibration.

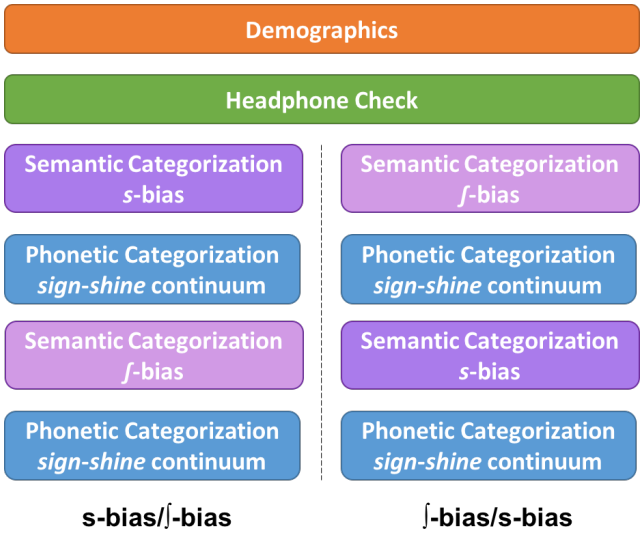

Figure 1. The general procedure for all experiments in this study.

Methods

Stimuli. Thirty-two words (16 with word-medial /s/, 16 with word-medial /ʃ/) were selected; the full set of stimuli is provided in Appendix A. 16 items (7 with medial /s/, 9 with medial /ʃ/) were taken directly from Kraljic and Samuel (2005), and the remaining items were generated following the same constraints that Kraljic and Samuel used to generate their stimuli. Student *t*-tests indicated no significant difference between /s/-medial and /ʃ/-medial words in written frequency (Kučera & Francis, 1967), $t(28) = 1.2$, $p = 0.24$, or in total number of syllables, $t(30) = 0.46$, $p = 0.65$. There was also no difference in the number of syllables preceding the critical fricative for /s/-medial words (mean: 1.69 syllables) and /ʃ/-medial words (mean: 2.00 syllables), $t(30) = -1.23$, $p = 0.23$.

As described in the Introduction, context that precedes the ambiguous token can guide perceptual learning, and it is unclear how much subsequent context can guide learning. In considering our stimuli, we noted that there may be variability in how strongly a particular phoneme is predicted by the phonemes that precede it. For instance, a stem like *epi-* has many

possible completions (*episode*, *epilogue*, *epiphenomenon*, etc.), but for a stem like *Arkan*-, there is only one possible completion in English (*Arkansas*). We thus opted to quantify how strongly the ambiguous speech sound could be predicted given the preceding phonemes in the word. If the preceding phonemes in the word are not perfectly diagnostic of the upcoming fricative, then a preceding sentence context might be able to provide a listener with stronger cues as to the identity of the ambiguous phoneme. To this end, we computed the frequency-weighted probability that the intended fricative would be the next phoneme given the preceding phonemes in the word. That is, for the word *episode*, we calculated the probability that the next phoneme would be /s/ given that the preceding phonemes were [ɛpɪ], accounting for the word frequency of each of the possible completions. Probabilities were calculated using the English Lexicon Project (Balota et al., 2007). We used the database to generate phonetic transcriptions for each word, and these transcriptions were then used to find all the words that began with the same onset as well as each onset competitor's written frequency (Kučera & Francis, 1967). In this way, we calculated how frequently the intended fricative occurred relative to all possible subsequent phonemes.[1] This analysis showed that the intended fricative (that is, the one that was replaced by an ambiguous phoneme) had a mean probability of 0.43 (SE: 0.07), which did not differ significantly between /s/-medial (0.43) and /ʃ/-medial (0.42) words, $t(28) = 0.04$, $p = 0.97$.

[1] Certainly, there may be some limitations to this particular analysis. For instance, it is not immediately apparent whether the relevant comparison should be the probability of the intended fricative compared to all possible continuations or the probability of the intended fricative relative to some subset of the possible continuations (e.g., all fricatives). Further, this metric ignores part of speech (which listeners may have strong predictions about, at least following predictive sentence contexts). Nonetheless, we present this as a coarse metric to show that the preceding phonemes are not necessarily diagnostic of the subsequent phoneme (i.e., the critical fricative does not necessarily occur after the word's uniqueness point).

We included roughly equal numbers of words that were concrete nouns and words that were not. Note that unlike a lexical decision task, answers for the semantic judgment task are rather subjective, as it is not immediately apparent whether some of our items (e.g., *Arkansas*) are concrete nouns or not. Based on experimenter judgment, however, approximately 14 items were concrete nouns and 18 were not; a chi-square test of independence indicated that status as a concrete noun was independent from whether words contained a medial /s/ or /ʃ/, $\chi^2(1) = 0.16$, $p = 0.69$.

Stimuli were produced by a female native speaker of American English, who was recorded in a sound-attenuated booth using a RØDE NT-1 condenser microphone with a Focusrite Scarlet 6i6 digital audio interface. The talker produced both a lexically consistent token (e.g., *episode*) as well as a lexically inconsistent nonword token (e.g., *epishode*) for each item; as described below, these tokens ultimately served as endpoints to generate word-nonword continua from which ambiguous tokens were selected. The talker produced each token (word and nonword) after each of its corresponding sentence contexts (see Experiment 2), with two productions recorded for each token. The speaker paused before each critical token to reduce the impact of coarticulation on the target item. Finally, the speaker also produced five productions each of the words *sign* and *shine* to generate stimuli for the phonetic categorization task.

Following recording, the default noise reduction filter was applied to the entire audio file in Audacity (Mazzoni & Dannenberg, 2015). Sentence-final tokens (e.g., *episode*, *epishode*) were cut at zero-crossings in Praat (Boersma & Weenik, 2017), and the first author then selected what he judged to be the best production of each lexically consistent item (*episode*) and each lexically inconsistent item (*epishode*). These tokens were scaled to a mean amplitude of 70 dB SPL. For each item, an 11-step word-nonword (e.g., *episode-epishode*) continuum was generated using

STRAIGHT (Kawahara et al., 2008), a software that allows for holistic morphing between two endpoint audio files; in STRAIGHT, the experimenter manually identifies landmark points in the endpoint stimuli (e.g., recordings of *episode* and *epishode*) in both the temporal and spectral domains prior to interpolation, resulting in a continuum that is more naturalistic than one that would be produced through waveform averaging alone; note that this procedure means that endpoint stimuli need not be the same duration prior to generating a continuum. An 11-step continuum was also generated from *sign* to *shine* to be used in the phonetic categorization post-test. Based on experimenter judgment, we decided that step 7 would provide a suitably ambiguous fricative for each continuum; note that the continuum was asymmetric, as step 7 was not the middle step along the generated continuum but was perceptually judged to be the most ambiguous. Step 4 was selected to serve as the clear /s/ token for each item, and step 10 was selected to serve as the clear /ʃ/ token. Thus, all fricative-containing tokens had been morphed in STRAIGHT, and endpoint tokens were an equal number of steps away from the ambiguous token. Similarly, steps 4-10 from the *sign-shine* continuum were selected for use in the phonetic categorization task.

Participants. Ninety-two participants were recruited for Experiment 1. Forty-three of these participants were recruited through the University of Connecticut's Psychology participant pool and completed the experiment in the lab. The other 49 participants were recruited using Amazon Mechanical Turk (MTurk), a crowdsourcing platform that has previously been used to study phonetic recalibration (Kleinschmidt & Jaeger, 2015; Liu & Jaeger, 2018). To qualify for the study, MTurk participants had to have the US set as their location and also had to have indicated that American English was the only language they spoke prior to age 13, that they had normal or corrected-to-normal hearing in both ears, and that their computer played sound.

All participants also completed a short auditory test designed to assess whether they were using headphones (Woods, Siegel, Traer, & McDermott, 2017). After an individual participated in one experiment, they were deemed ineligible to participate in subsequent experiments reported in this paper. In-lab participants received course credit for their participation. MTurk participants were paid $5.05 for completing the full experiment and $0.85 if they were deemed ineligible after completing the initial demographics screener. Payment amounts were based on estimated maximum time to complete the full experiment and the screening, respectively, multiplied by Connecticut's minimum wage of $10.10 per hour (at the time the study was conducted).

Participants were excluded from analyses if they failed to respond to a substantial portion (≥ 10%) of the trials on either task and/or if they showed poor accuracy (≤ 70%) in phonetic categorization of the unambiguous endpoints, similar to procedures followed in other web-based phonetic recalibration studies (e.g., Kleinschmidt & Jaeger, 2015). In this way, data from 12 participants (3 in-lab, 9 MTurk) were excluded. As such, the total sample size for Experiment 1 was 80 participants (44 women), with 40 subjects having participated in-lab and 40 having participated via MTurk.

Procedure. The full procedure is summarized in Figure 1. Following the demographics screener and the headphone check, eligible participants completed four experimental blocks. In the first and third blocks, participants completed a semantic categorization task. One block was /s/-biased and the other was /ʃ/-biased; block order (/s/-biased or /ʃ/-biased) was counterbalanced across participants. For these blocks, participants were told that they would hear a word on every trial and would need to decide if it was a concrete noun. Participants were asked to respond as quickly as possible. A concrete noun was defined in the instructions as a person, place or thing that can be experienced with any of the five senses (sight, sound, smell, touch, taste). In the /s/-

biased block, participants heard the ambiguous fricative only in contexts where lexical information disambiguated the sound as a /s/ (e.g., *epi?ode*) and also heard clear /ʃ/ endpoints in lexically congruent contexts (e.g., *friendship)*. In the /ʃ/-biased block, participants heard the ambiguous fricative only in /ʃ/-biasing conditions (*friend?ip*) and a clear /s/ in lexically congruent contexts (*episode*). Item order was randomized for each participant, and each participant heard all 32 items (16 /s/-biased, 16 /ʃ/-biased) each time they completed the semantic categorization task. There were no filler items.

During the second and fourth blocks of the experiment, participants completed a phonetic categorization task. They were told they would hear the word *sign* or *shine* on each trial and to indicate as quickly as possible which one they heard. Participants heard each step from the 7-step continuum ten times presented in a random order, yielding a total of 70 trials for each block.

For both tasks, participants were prompted to indicate their response with the keyboard after they heard the stimulus; button mappings were counterbalanced across participants. Participants had 4 seconds to make their response before the trial timed out, and there was a 1-second interval between trials. The experiment was programmed using custom JavaScript code using functions from the jsPsych library (de Leeuw, 2015) and was hosted online using Google App Engine. All stimuli, code and outputs from statistical models are publicly available at https://osf.io/eqwja/ (Luthra, Magnuson, & Myers, 2020). The full session took approximately 30 minutes, and procedures for all the experiments reported in this paper were approved by the University of Connecticut's Institutional Review Board. In all experiments, subjects gave informed consent prior to participating.

Results

Results from the phonetic categorization task, in which participants categorized items along a *sign-shine* continuum, are plotted in Figure 2. Recalibration is apparent as a difference in the pattern of phonetic categorization following /ʃ/-biased context, shown in green (dark) lines, as compared to the /s/-biased contexts, shown in orange (light) lines.

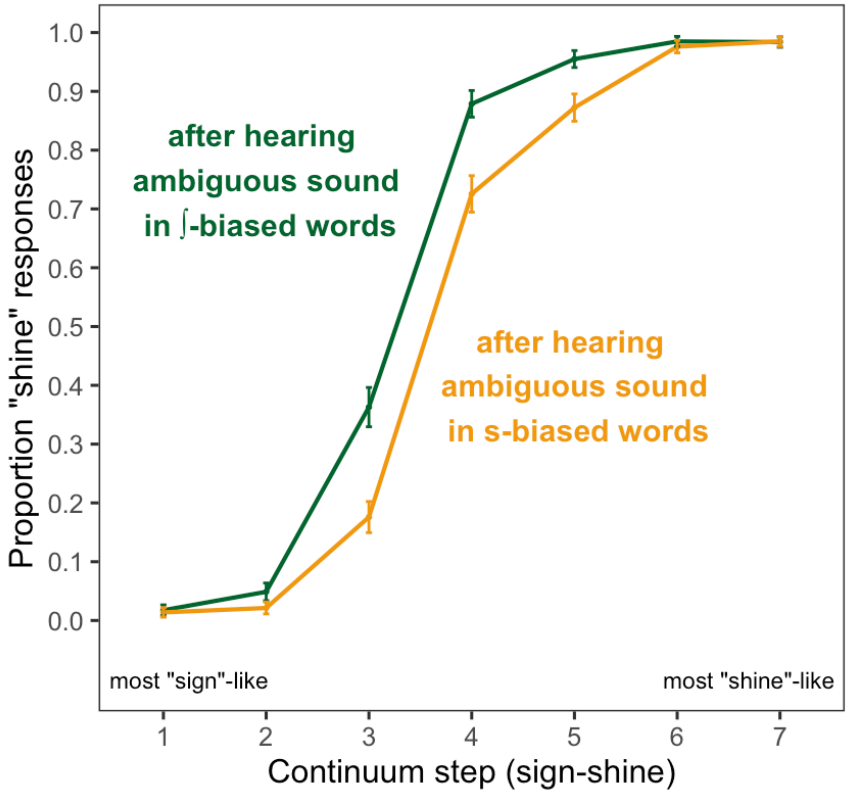

Figure 2. Data from the phonetic categorization task of Experiment 1, in which all words were presented in isolation, without any sentence context. Error bars indicate 95% confidence intervals.

Data were analyzed using mixed effects logistic regression in R (R Core Team, 2018). Models were implemented the *lmer* function from the "lme4" package (Bates, Maechler, Bolker, & Walker, 2015); main effects and interactions were estimated via likelihood ratio tests using the

mixed function in the "afex" package (Singmann, Bolker, Westfall, & Aust, 2018). In Experiment 1, a mixed model considered fixed factors of Bias (s-bias, ʃ-bias; coded with a [-1, 1] contrast) and Step (centered). For all analyses, random effect structures were determined as follows. Following the recommendation of Barr, Levy, Scheepers and Tily (2013), we began with the maximal random effect structure (i.e., one that included by-subject random intercepts and random by-subject slopes—as well as their interactions—for all factors manipulated within-subject). If the maximal model did not converge, simpler random effect structures were tested; first, random correlations were removed, followed by random slopes (with random effects of Step removed before random effects of Bias) and finally random interactions. Once convergence was achieved, a backward stepping procedure was used to determine whether the random effect structure could be further simplified without a significant loss in model fit, as assessed by a chi-square test; by using this stepping procedure, we sought to ensure that we were using a parsimonious model structure, preventing against Type I error without sacrificing power to detect fixed effects (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). For Experiment 1, this procedure identified the maximal model as the best random effect structure; this included random by-subject intercepts, random by-subject slopes for Bias and Step, and random by-subject interactions between Bias and Step.

In Experiment 1, we observed a significant effect of Bias, $\chi^2(1) = 42.51$, $p < 0.001$, consistent with phonetic recalibration. We also observed an expected effect of Step, $\chi^2(1) = 162.56$, $p < 0.001$, indicating that participants made more /ʃ/ responses as the continuum tokens became more /ʃ/-like. The interaction between Bias and Step approached significance, $\chi^2(1) = 3.46$, $p = 0.06$, indicating that the effect of Bias may not have been constant at all steps.

Because half of the data came from in-lab participants and half came from MTurk participants, we also examined whether there were differences in the size of the perceptual learning

effect across settings. As reported in Appendix B (Table B1), there were no significant interactions between Bias and Setting.

As discussed in the Introduction, Bias is not typically manipulated within subjects, since previous work has shown that phonetic recalibration does not occur if listeners hear clear productions of a fricative prior to hearing ambiguous productions (Kraljic, Brennan, et al., 2008; Kraljic & Samuel, 2011; Kraljic, Samuel, et al., 2008). In the current study, however, we were able to successfully manipulate Bias within subjects in the current experiment, as has been done in other phonetic recalibration studies that have not used fricative contrasts (Bonte et al., 2017; Kilian-Hütten et al., 2011; van Linden & Vroomen, 2007). To address concerns that hearing both biasing conditions may have affected our results, we conducted two additional analyses. First, we performed an analysis where we only examined data from the first phonetic categorization block (effectively making Bias a between-subjects factor). For this analysis, random by-subject slopes for and interactions with Bias were dropped, as subjects were only exposed to one biasing condition prior to this block; otherwise, the model structure was the same as in the main analysis. We observed the same set of results as in the main analysis – namely, a significant effect of Bias, $\chi^2(1) = 12.87, p < 0.001$, a significant effect of Step, $\chi^2(1) = 148.84, p < 0.001$, and a non-significant interaction between Bias and Step, $\chi^2(1) = 0.50$, $p = 0.48$. In a second analysis, we examined whether effects of Bias were stable across blocks. As reported in Appendix B (Table B2), there were no interactions between Bias and Block.

Discussion

The results of Experiment 1 indicated that phonetic recalibration occurred when participants were given a semantic categorization task during exposure, consistent with previous

studies that have demonstrated LGPL effects without using a lexical decision task (Drouin & Theodore, 2018; Eisner & McQueen, 2006; Leach & Samuel, 2007; Maye et al., 2008; McQueen et al., 2006). To our knowledge, this is the first LGPL study to use a concreteness judgment during exposure. In another departure from the standard paradigm, we manipulated Bias within subjects (following van Linden & Vroomen, 2007); the fact that we observed phonetic recalibration even after participants had heard clear productions of the fricative sounds is inconsistent with previous findings on phonetic recalibration (e.g., Kraljic, Samuel, et al., 2008) and suggests that listeners' ability to override previous experience with a talker's voice may be greater than has been previously described in the literature. Notably for the current study, the ability to manipulate Bias within subjects allows us to minimize the influence of subject-to-subject variability on our measurement of the Bias effect, thus affording us more power to detect interactions between Bias and Context in subsequent experiments (Experiments 2-4).

Having ascertained that our paradigm can successfully induce phonetic recalibration, we turn next to the experiments designed to examine whether boosting lexical support through a preceding context can modulate the size of perceptual learning effects.

Experiment 2

In Experiment 2, we examined whether the extent of LGPL can be affected by whether a preceding sentence context predicts the identity of a word containing an ambiguous segment. One group of participants heard a predictive auditory context (e.g., *I love "The Walking Dead" and eagerly await every new…*) before each target item (e.g., *episode*), while another group heard neutral sentence contexts (*My ballpoint pen ran out of ink when I was halfway through writing the word…*). We expected that Context (predictive / neutral) would modulate the size of the Bias (s-

bias / ʃ-bias) effect (i.e., we expected a Bias × Context interaction). As discussed earlier, our key question is whether the recalibration effect would be larger for predictive contexts, as would be predicted by an ideal observer account (Kleinschmidt & Jaeger, 2015), or whether predictive sentential context would attenuate learning by shifting attention away from phonetic detail, as would be predicted by an attentional weighting account (Goldstone, 1998).

Methods

Stimuli. We used the 32 words described in the Methods for Experiment 1 (16 with word-medial /s/, 16 with word-medial /ʃ/). For each item, we created two predictive contexts and two neutral contexts. Two contexts were needed per item because every subject was exposed to each item twice (once in the /s/-biased exposure block and once in the /ʃ/-biased block), and we did not want subjects who were receiving neutral contexts to be able to predict the sentence-final target during their second exposure block (on the basis of their memory for sentence contexts from the first exposure block). As such, we created one set of sentence contexts for the first exposure block and a separate set of contexts for the second exposure block.

The predictive power of our sentence contexts was assessed with a norming pretest. In the pretest, participants were given sentence contexts and asked to complete each one with the first word that came to mind (the cloze procedure; Taylor, 1953). Each participant saw only one of the two sentences that were designed to predict a particular item. Occasionally, some participants withdrew before completing all sentences, so a total of 65 participants were recruited in order to collect 20 responses for each sentence context. Participants were recruited through Amazon Mechanical Turk and compensated at a rate of $10.10/hour. The cloze probability of the target word in each predictive sentence context is listed in Appendix A. The intended target had a mean

cloze probability of 0.74 in predictive contexts (SE: 0.03), and this did not differ between /s/-medial and /ʃ/-medial targets, $t(62) = 0.11$, $p = 0.91$. Cloze probability ratings did not differ between the predictive contexts that appeared in the first exposure block and the contexts that appeared in the second block, $t(31) = 1.61$, $p = 0.12$. Neutral contexts never elicited their associated target items.

Sentence contexts did not include /s/ or /ʃ/ phonemes.[2] However, sentence contexts did include other fricatives (/f/, /v/, /θ/, /ð/, /z/, /ʒ/, /h/), as excluding those phonemes would have dramatically limited the scope of possible words; notably, this includes the voiced versions of the critical segments /s/ and /ʃ/ (/z/ and /ʒ/, respectively). These two fricatives occurred 146 times across the 128 sentence contexts, and a chi-square test of independence indicated that these fricatives occurred with roughly equal likelihood across predictive and neutral contexts as well as across /s/- and /ʃ/-biased contexts, $\chi^2(1) = 0.58$, $p = 0.45$. In Experiment 3, we consider the possibility that hearing unaltered productions of these fricatives could affect phonetic recalibration for /s/ and /ʃ/.

Sentence contexts had a mean length of 14.5 words. An analysis of variance indicated that there were no differences in sentence length as a function of the target's medial fricative (/s/ or /ʃ/), $F(1,30) = 0.18$, $p = 0.68$, as a function of the type of context (neutral or predictive), $F(1,30) = 1.18$, $p = 0.29$, or as a function of whether the sentence context would be used in the first or second exposure block, $F(1,30) = 1.77$, $p = 0.19$. There were also no significant interactions between any of these factors.

[2] During recording, it was noted that three normed contexts each contained an instance of /s/, so we opted to record minimally altered sentences. While these new contexts were not identical to the ones normed, we do not expect these minimal changes to substantially affect cloze probabilities. In particular, "on the first day of camp" was changed to "at the beginning of camp;" the word "interesting" was changed to "intriguing;" and an instance of "so" was changed to "and."

Sentence contexts were recorded during the same recording session as critical target items (see Methods for Experiment 1). Sentence contexts were excised from the auditory file by cutting at zero-crossings in Praat; the first author then selected the context he deemed to be the best recording. These contexts were then scaled to a mean amplitude of 72 dB in Praat and concatenated with the sentence-final words (which had been scaled to a mean amplitude 70 dB). In this way, the recordings of the critical items (e.g., *epi?ode*) were the same as the recordings used in Experiment 1. Note that we used different dB values for the sentence context and the critical item to equate for differences in perceived amplitude.

Participants. Data from 177 participants were collected for Experiment 2, with 83 participants recruited from the University of Connecticut's Psychology participant pool and 94 participants recruited via MTurk. As before, we excluded the data of participants who failed to respond to at least 10% of the trials on either the semantic categorization or phonetic categorization tasks as well as the data from participants whose categorization of the continuum endpoints was at or below 70%. This resulted in the exclusion of 17 participants (3 in-person, 14 from MTurk), yielding a total sample size of 160 participants (96 women). Of these, 80 participants completed the experiment in person (40 receiving predictive contexts, 40 receiving neutral contexts), and 80 completed the experiment via MTurk (with half receiving each type of sentence context).

Procedure. The procedure for Experiment 2 was identical to that followed for Experiment 1, with the exception that the exposure trials involved the presentation of an auditory sentence context prior to the critical word. Participants were told that their task during the exposure blocks was to decide if the final word of each sentence was a concrete noun.

Results

Exposure. In a previous semantic priming study, McRae and Boisvert (1998) demonstrated that participants were faster to decide if a written target was a concrete noun if the target was preceded by a semantically related prime. As such, we tested whether concreteness judgments made during exposure were faster when participants received a predictive sentence context compared to a neutral one. Reaction time data were measured from word offset. Following the approach of Lo and Andrews (2015), we employed a mixed effects model that included a gamma distribution and an identity link function. This model considered whether response times differed based on Context (neutral, predictive; coded with a [-1, 1] contrast). The model included only by-subject random intercepts; as Context was manipulated between subjects, this was both the maximal and most parsimonious random effects structure. We observed a significant effect of Context, $\chi^2(1) = 4.92$, $p = 0.03$, driven by faster responses after predictive contexts (mean: 610 ms, SE: 8 ms) than after neutral contexts (mean: 713 ms, SE: 9 ms).

Phonetic categorization. Results from Experiment 2 are plotted in Figure 3. As before, green (dark) lines indicate phonetic categorization responses after /ʃ/-biasing blocks, while orange (light) lines indicate categorization after /s/-biasing blocks; as such, the difference between lines of different colors indicates the size of the phonetic recalibration effect. Evidence for an influence of sentential context on the size of recalibration effects would manifest as a difference in the size of the phonetic recalibration effect after predictive sentence contexts (shown in solid lines) as compared to neutral contexts (shown in dashed lines). The trends apparent in the figure are for sizeable differences between fricative bias conditions but no apparent impact of sentence context.

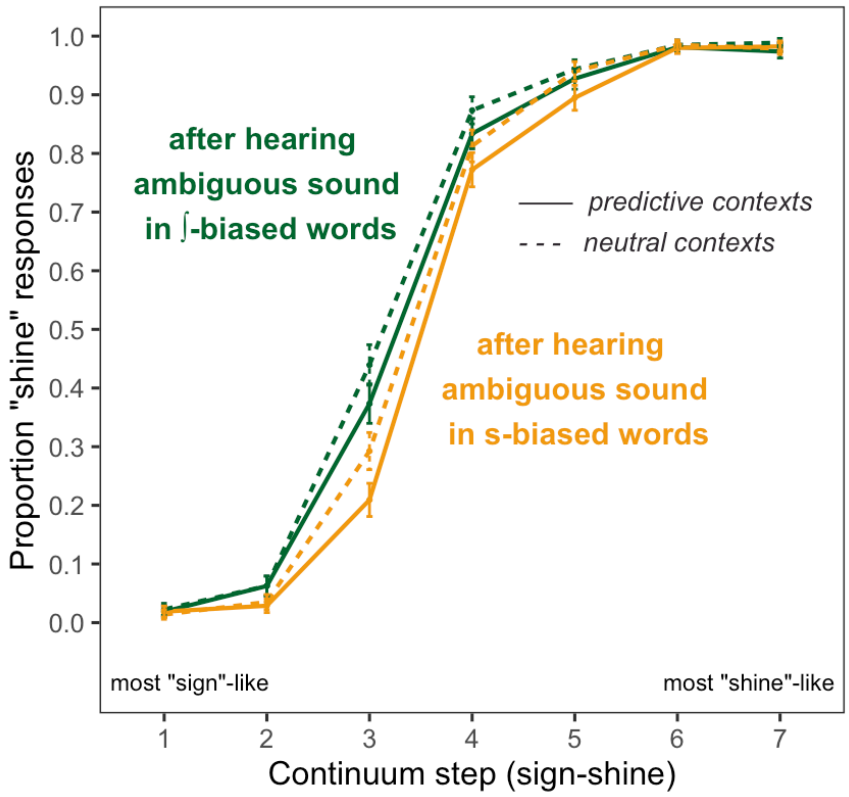

Figure 3. Data from the phonetic categorization task in Experiment 2, in which participants heard neutral or predictive sentence contexts prior to items with word-medial fricatives. Error bars indicate 95% confidence intervals.

Data from Experiment 2 were analyzed using a mixed effects logistic regression model that considered fixed factors of Bias (s-bias, ∫-bias; coded with a [-1, 1] contrast), Step (centered), and Context (Neutral, Predictive; coded with a [-1, 1] contrast); random intercepts for each subject were also modeled. Results are given in Table 1. The model indicated the expected effect of Bias, $p < 0.001$, consistent with phonetic recalibration, and the expected effect of Step, $p < 0.001$. However, the expected interaction between Bias and Context was not observed, $p = 0.79$, suggesting effects were not different between subjects receiving predictive contexts and those who

received neutral contexts. There was a marginal effect of Context, $p = 0.05$, but no other significant results were observed.

Effect	**χ2(1)**	***p***
Bias	18.42	< 0.001
Step	324.69	< 0.001
Context	3.72	0.05
Bias × Step	2.56	0.11
Bias × Context	0.07	0.79
Step × Context	1.21	0.27
Bias × Step × Context	0.00	0.95

Table 1. Phonetic categorization results from Experiment 2.

As in Experiment 1, half of the participants were recruited through MTurk and half participated in an in-lab study. A supplementary analysis (Appendix B, Table B3) indicates that the size of perceptual learning effects did not differ across samples.

Also as in Experiment 1, participants were exposed to both /s/- and /ʃ/-biasing conditions, with order counterbalanced. Because Bias is not typically manipulated within subjects, we also conducted an analysis that only considered data from the first block of phonetic categorization (effectively making Bias a between-subjects factor). We obtained a similar pattern of results with this analysis (Appendix B, Table B4) as compared to our main analysis (Table 1). As before, we also performed an analysis with Block as a factor to examine whether effects of Bias were comparable across blocks; we observed no interactions between Bias and Block (Appendix B, Table B5), indicating a comparable extent of recalibration in each block.

Discussion

In Experiment 2, we observed phonetic recalibration when the critical stimuli were presented in an auditory sentence context, consistent with other LGPL studies where critical items were embedded in sentence contexts (Eisner & McQueen, 2006; Maye et al., 2008). However, we did not observe an interaction between Context and Bias, suggesting that the size of the LGPL effect was not modulated by whether the sentence context was predictive of the critical word or neutral with respect to it. This finding is contrary not only to the predictions of an ideal observer account, which predicts larger recalibration effects for predictive contexts compared to neutral ones (Kleinschmidt & Jaeger, 2015), but also contrary to the predictions of an attentional weighting account (Goldstone, 1998), which predicts larger recalibration effects in neutral contexts than predictive ones.

Notably, context did have a strong influence on online processing: Participants who heard predictive contexts were significantly faster in making their semantic categorization decisions than were participants who heard neutral contexts. However, this did not translate to a difference in the size of perceptual learning effects. These findings suggest that sentence context may not have a strong influence on how lexical information guides phonetic learning.

Strikingly, we found in both Experiments 1 and 2 that participants could learn to recalibrate to one talker but then re-adjust their perceptual boundaries on the basis of new, contradictory evidence about the talker's speech. This suggests that listeners may be more flexible in adapting to talker-specific phonetic variation than has been previously described (Kraljic, Samuel, & Brennan, 2008); we return to this point in the General Discussion.

It is also noteworthy that though the sentence contexts we created did not include other instances of /s/ or /ʃ/, these contexts did include other fricatives, including voiced versions of /s/ and /ʃ/ (/z/ and /ʒ/, respectively). It is possible that the perceptual units that are recalibrated during

LGPL are sub-phonemic units; that is, rather than learning how a talker produces individual phonemes, listeners may be learning how a talker produces particular acoustic-phonetic features, such that learning should apply to larger classes of speech sounds (e.g., all fricative sounds; Kraljic & Samuel, 2006). As such, auditory exposure to other fricatives during the sentence contexts may have diminished learning and thus may have potentially obscured group differences in learning effects. In Experiment 3, we addressed this concern by presenting critical items after *written* sentence contexts, which do not provide additional auditory information about the talker.

Experiment 3

In Experiment 3, exposure blocks consisted of auditory stimuli with word-medial fricatives that were presented after written sentence contexts. The use of written contexts allows us to provide higher-level context for the target items but without providing subjects with auditory exposure to the fricatives (e.g., /z/, /ʒ/) in the sentence contexts, which could conceivably disrupt the impact of our critical /s/-/ʃ/ items. As in Experiments 1 and 2, exposure blocks were followed by phonetic categorization blocks. As in Experiment 2, we were specifically interested in whether we would observe larger recalibration effects with predictive contexts or with neutral ones.

Methods

Stimuli. We used the same stimuli as in Experiments 1 and 2.

Participants. 206 participants were recruited for Experiment 3. For this experiment, all participants were recruited through MTurk; note that in Experiments 1 and 2, we did not observe any differences in the size of learning effects between participants recruited through MTurk and participants recruited through the University of Connecticut participant pool. We followed the

same exclusion criteria as in previous experiments, resulting in discarding the data of 46 participants. Of the 160 participants (63 women) included in analyses, 80 received predictive contexts and 80 received neutral ones.

Procedure. We followed the same procedure as in Experiment 2 with one important exception: Instead of presenting sentence contexts in the auditory modality during exposure blocks, written contexts were provided. To encourage participants to read the full context, we used a self-paced reading design, where participants were only shown one word of the sentence at a time and had to press the spacebar to see the next word. Text was presented at twice the browser's default font size (2em) in center-aligned black Open Sans text on a white background. The final word of the sentence was presented in the auditory modality, and participants made the same semantic categorization judgment as in the previous experiments, with stimulus timings and button mappings set as in Experiment 1.

Results

Exposure. As before, we first examined whether response times during the semantic categorization task were influenced by the type of context a participant had received. A mixed effects analysis (implemented identically to the model for Experiment 2) indicated a significant effect of Context, $\chi^2(1) = 4.28$, $p = 0.04$, driven by significantly faster response times after predictive contexts (mean: 616 ms, SE: 8 ms) than after neutral ones (mean: 690 ms, SE: 8 ms).

Phonetic categorization. Results from the phonetic categorization task are shown in Figure 4. Though phonetic recalibration was robustly observed (as illustrated by the difference between dark and light lines), the size of the recalibration effect does not appear to differ based on whether participants received predictive contexts (solid lines) or neutral contexts (dashed lines).

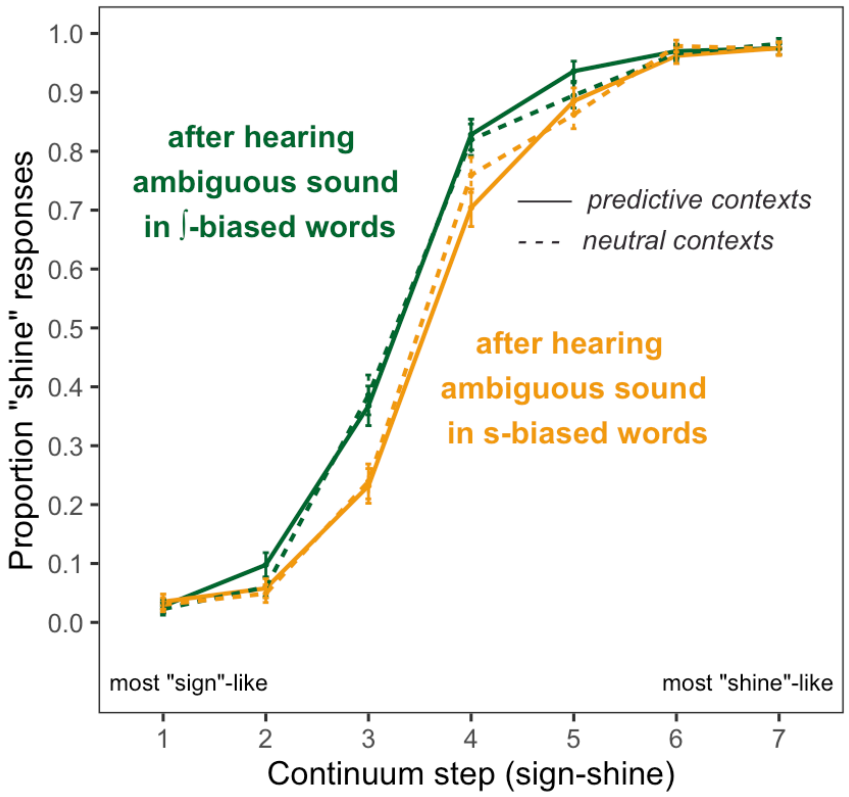

Figure 4. Data from the phonetic categorization task in Experiment 3, in which participants read neutral or predictive sentence contexts prior to items with word-medial fricatives. Error bars indicate 95% confidence intervals.

Phonetic categorization data were modeled following the same approach as in Experiment 2. Results are provided in Table 2. We observed the expected effect of Bias, $p < 0.001$, demonstrating that phonetic recalibration had occurred, as well as an effect of Step, $p < 0.001$. However, no other effects were significant.

Effect	**χ2(1)**	***p***
Bias	27.69	< 0.001
Step	285.68	< 0.001
Context	0.48	0.49

Bias × Step	0.11	0.74
Bias × Context	1	0.32
Step × Context	0.94	0.33
Bias × Step × Context	0.19	0.67

Table 2. Phonetic categorization results from Experiment 3.

As in previous experiments, we also conducted an analysis that only considered the first block of phonetic categorization data, such that Bias was rendered a between-subjects factor. Results (Appendix B, Table B6) largely resembled the results presented in the main text (Table 2). Furthermore, we conducted an analysis that tested whether Bias effects differed across block; results (Appendix B, Table B7) suggest stable effects of Bias throughout the experiment.

Discussion

To test whether the lack of Bias × Context effects in Experiment 2 arose because of auditory exposure to other fricatives in the sentence contexts, Experiment 3 used written sentence contexts prior to each auditory stimulus. In this way, participants received the contextual support of the sentence context without the additional auditory information that might have attenuated learning in Experiment 2. While there are important differences between auditory contexts and self-paced reading, including the fact that the context is presented at a variable rate in the self-paced reading task, results indicate that context influences online processing of the sentence-final target items (an effect of Context in the semantic categorization task) and that phonetic recalibration occurs with this paradigm (evidenced by an effect of Bias in the phonetic categorization task), suggesting that this paradigm is indeed an appropriate one for the present investigation.

While the type of context participants received did have an influence on in-the-moment processing of concreteness information (as evidenced by a main effect of Context in the semantic

categorization data), the type of context did not influence the degree of learning in the phonetic categorization phase (as evidenced by a non-significant Bias × Context interaction). Taken together, the results of Experiments 2 and 3 suggest that the extent of lexically guided perceptual learning is not affected how strongly the critical lexical items are predicted by preceding sentence contexts.

Experiment 4

In Experiments 2 and 3, participants received sentential contexts that were either predictive of the upcoming critical word or neutral with respect to it, thereby manipulating how strongly the listener expected a particular target item. The ultimate goal of these manipulations was to vary the degree of preceding lexical support for the critical target word. While norming data indicate that the predictive sentence contexts elicited the intended target word significantly more often than did the neutral contexts, there was considerable variability in the predictive sentence contexts with regard to the cloze probabilities for the target word. In Experiment 4, we manipulated the degree of lexical support more directly by either showing participants the written form of the target word (e.g., *episode*) prior to hearing it or by showing them filler text (#########); such an approach has been shown to have strong effects in a variety of paradigms (e.g., Blank & Davis, 2016). Because written words were always followed by the identical auditory word, this manipulation provided the strongest possible cue to the identity of the upcoming word.

Methods

Stimuli. We used the auditory stimuli as in previous experiments.

Participants. 215 people were recruited through MTurk for Experiment 4. Following our established inclusion criteria, we discarded data from 55 participants, resulting in a sample size of

160 participants (68 women), of whom 80 received predictive contexts and 80 received neutral contexts.

Procedure. The procedure for this experiment resembled that of Experiment 3 with one important change to the semantic categorization task. Rather than reading a sentence context prior to hearing the auditory token, participants in the Predictive group saw the written form of the word that they were going to hear, whereas participants in the Neutral group saw control text (########). Once the participant pressed the spacebar, they heard the auditory target and made their semantic categorization.

Results

Exposure. During the exposure block, participants could choose how long they spent reading the text that preceded each auditory target. As such, we examined how long participants read the text that preceded words with an ambiguous fricative (e.g., *epi?ode*, *friend?ip*) compared to words with an unambiguous fricative (*episode*, *friendship*). For this analysis, we excluded trials with a reading time of more than 6 seconds, resulting in the exclusion of 169 trials (1.65% of all data). Reading time data were analyzed using a generalized mixed effects model with a gamma distribution and an identity link function. We tested for fixed effects of Context (neutral, predictive; coded with a [-1, 1] contrast) and of fricative Ambiguity (unambiguous, ambiguous; coded with a [-1, 1] contrast); the model also included random by-subject intercepts and slopes for Ambiguity. We observed a significant effect of Context on reading times, $\chi^2(1) = 4.92$, $p = 0.03$, driven by longer reading times for predictive text (mean: 642 ms, SE: 9 ms) than for neutral text (mean: 543 ms, SE: 9 ms). However, there was no significant effect of Ambiguity on reading times, $\chi^2(1) = 0.46$, $p = 0.50$, nor an interaction between Context and Ambiguity, $\chi^2(1) = 2.32$, $p =$

0.13. Thus, while participants read predictive text for longer than neutral text, reading times were not influenced by whether the subsequent word contained an ambiguous fricative.

Semantic categorization data from the exposure task were analyzed using the same procedure as in previous experiments. We observed a significant effect of Context on response times, $\chi^2(1) = 29.83$, $p < 0.001$, driven by faster response times when participants received predictive contexts (mean: 447 ms, SE: 7 ms) than when they received neutral ones (mean: 666 ms, SE: 8 ms).

Phonetic categorization. Data from the phonetic categorization task are illustrated in Figure 5. Visually, it is clear that phonetic recalibration occurred (evidenced by the difference between the dark and light lines), but there are no obvious differences in the size of the recalibration effect as a function of the type of preceding context.

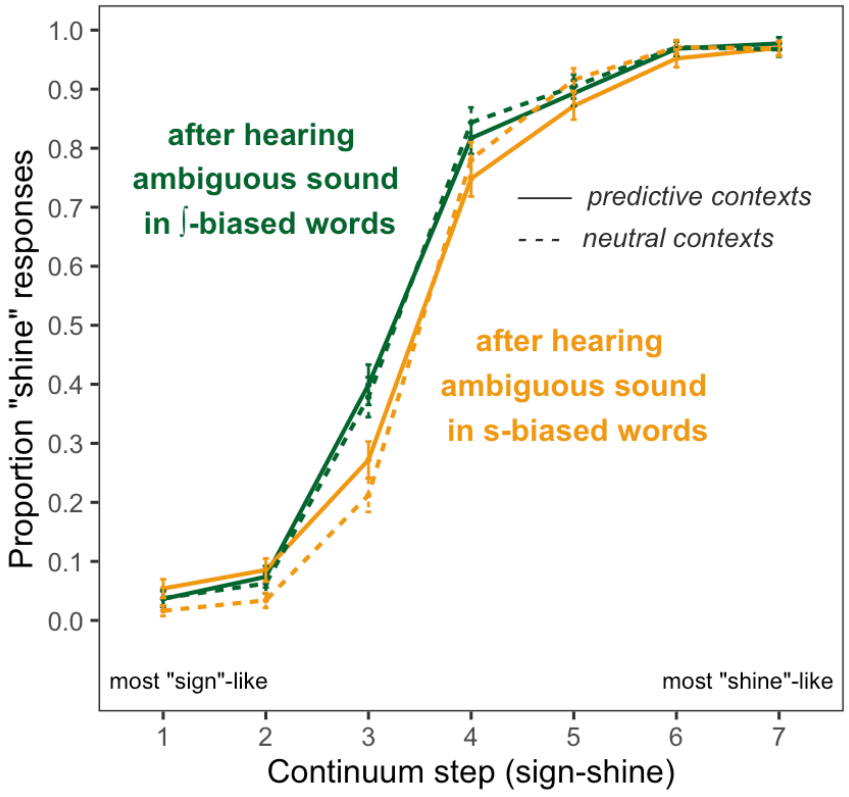

Figure 5. Data from the phonetic categorization task in Experiment 4, in which subjects either read the target word prior to hearing it (predictive condition) or read filler text *(########;* neural condition) before hearing the auditory stimulus. Error bars indicate 95% confidence intervals.

Data from the phonetic categorization task were analyzed using a logit mixed effects model, as before. As indicated in Table 3, we observed the expected effect of Bias, $p < 0.001$, indicating phonetic recalibration, as well as the effected expect of Step, $p < 0.001$, indicating perceptual sensitivity to the /s/-/ʃ/ manipulation. We also observed a significant Bias × Step interaction, $p = 0.03$, indicating that the size of the Bias effect was not constant across steps. We observed a marginal Step × Context interaction, $p = 0.05$, as well as a significant Bias × Step × Context interaction, $p = 0.03$. The latter interaction reflects the fact that at the "sign" end of the continuum, there was a slightly larger learning effect for the neutral context than for the predictive context, but at the "shine" end of the continuum, there was a slightly larger learning effect for the predictive context than for the neutral one.

Effect	**$\chi 2(1)$**	***p***
Bias	15.25	< 0.001
Step	289.33	< 0.001
Context	0.55	0.46
Bias × Step	4.69	0.03
Bias × Context	0.08	0.78
Step × Context	3.89	0.05
Bias × Step × Context	4.55	0.03

Table 3. Phonetic categorization results from Experiment 4.

As before, we performed an additional analysis that only considered the first phonetic categorization block (Appendix B, Table B8). This analysis yielded a similar pattern of results as described in Table 3 – namely, significant effects of Step, Step × Context, and Bias × Step × Context. We also performed an analysis testing for interactions between Bias and Block (Appendix B, Table B9) and found no evidence that the effect of Bias differed in size across blocks.

Discussion

In Experiment 4, auditory targets were preceded by either predictive text – specifically, the written form of the upcoming target word – or neutral text (########). We observed a robust effect of preceding context on semantic categorization judgments, suggesting that listeners were able to leverage context to guide in-the-moment processing. At test, we observed a Bias × Step × Context interaction, which was driven by larger learning effects in the predictive group on one side of the continuum and larger learning effects in the neutral group on the other side of the continuum. Such a pattern of results is not predicted by either theoretical account and is likely driven by noise in the data. That is, there do not appear to be consistent differences in the size of perceptual learning effects across the continuum, such that one group showed larger learning than the other. In summary, even when participants had very strong or very weak expectations about what word they were likely to see next, there were no theoretically meaningful differences in the size of the perceptual learning elicited by that word.

As described in the Introduction, previous work has shown that phonetic recalibration can be induced by text. In a study by Keetels et al. (2016), subjects were exposed to a segment of speech (a?a) containing a fricative that was ambiguous between /b/ and /d/, and simultaneously presented written text (*aba* or *ada*) guided phonetic retuning. Notably, the written text was the

only information participants in that study had to disambiguate the critical speech segment. By contrast, participants in Experiment 4 received written contexts prior to the auditory segment. Furthermore, the written information in Experiment 4 was not the only cue to the ambiguous segment's identity, as participants received all the information they needed about the segment's identity from lexical context (e.g., the knowledge that *episode* is a word and *epishode* is not). Thus, while previous work suggests that written context can in principle guide recalibration, the results of Experiment 4 suggest that preceding written context does not influence how lexical information guides perceptual learning. That is, a redundant written context does not have an additional influence on perceptual learning, over and above the influence of lexical information.

General Discussion

The current study was designed to examine how the predictive power of a preceding context might affect LGPL, as existing theoretical accounts make opposing predictions of how the degree of lexical support will influence the extent of recalibration. In a series of experiments using a modified LGPL paradigm, listeners were exposed to a talker who produced a fricative that was ambiguous between /s/ and /ʃ/ in lexical contexts that biased the listener to interpret the ambiguity in a particular way. While some participants heard the critical words without any preceding context (Experiment 1), others received auditory (Experiment 2) or written (Experiments 3 and 4) contexts that were either predictive of or neutral with respect to the upcoming critical word.

Importantly, phonetic recalibration was observed across all experiments, even though our paradigm deviated from the standard paradigm in two key ways. Firstly, we used a semantic categorization task during exposure, rather than the standard lexical decision task; as such, we have expanded the set of paradigms that can successfully drive LGPL. Secondly, we manipulated

Bias within subjects. That is, listeners were first exposed to a talker for whom ambiguous tokens were biased to be interpreted as one phoneme (e.g., /s/); after testing for recalibration, listeners were exposed to the same talker, with ambiguous tokens now mapping to the other fricative category (e.g., /ʃ/). Effects of Bias were robustly observed across all four experiments, with listeners demonstrating a consistent ability to update the mapping between acoustics and phonetic categories on the basis of new information about a talker, even when that information conflicted with previous exposure. This suggests that previous work (Kraljic, Samuel, & Brennan, 2008) may have underestimated just how flexible listeners are in phonetic retuning and underscores the importance for additional work examining how listeners deal with conflicting information during the course of phonetic recalibration.

Preceding context had a clear influence on listeners' in-the-moment processing of auditory stimuli: Across all studies that included preceding context (Experiments 2-4), predictive context facilitated concreteness judgments on auditory targets encountered immediately after the context. However, context did not influence the extent of perceptual learning, as it was never the case that one group showed larger learning effects than the other across the continuum. One possibility is that phonetic recalibration is an all-or-nothing phenomenon, and it may not be possible to observe differences in the size of learning effects as a function of some group manipulation. Alternatively, it is possible that learning differences might emerge under some conditions, but simply failed to emerge in the specific conditions we tested.

In considering why we failed to see interactions between the nature of preceding context and the size of learning effects, it is important to remember that the goal of the current study was to test whether the strength of learning could be affected by how strongly a preceding context predicted an upcoming target word. As such, the identity of the ambiguous phonemes in our

experiment could be resolved simply based on a listener's lexical knowledge (i.e., knowing that *episode* is a word). Previous research has established that if lexical context does not disambiguate a talker's intended phonetic category (i.e., with minimal pair items like "goat" and "coat"), sentential context can guide a listener's in-the-moment interpretations of phonetically ambiguous speech (Borsky, Tuller, & Shapiro, 1998; Guediche, Salvata, & Blumstein, 2013) and can guide perceptual learning for speech (Jesse, 2020). One interpretation of the present results is that when lexical information is enough to resolve the identity of an ambiguous phoneme, listeners may not strongly leverage sentence-level context to guide learning. Alternatively, it is possible that listeners do leverage sentence-level information but that the difference in the predictive nature of the context may no longer matter after listeners have encounter the critical word. That is, it may be the case that when listeners encountered predictive contexts, they received early information about the upcoming target word and received relatively little new information when they ultimately encountered the target word itself. By contrast, when participants heard neutral contexts, the target word was ultimately more informative. Thus, listeners may have ultimately received the same amount of lexical support, even if they received this support at different time points (i.e., relatively distributed throughout the preceding context or more concentrated at the time when they heard the final word), and thus the degree of perceptual learning may be identical across forms of context.

There are, of course, other reasons why we may have failed to see strong influences of sentence context in the current set of experiments. Notably, our experiments employed conditions that favored robust learning, and with a relatively large degree of learning, it might be hard to observe robust effects of sentence context. This sort of a ceiling effect might have emerged in part because in the current study, listeners never encountered nonwords, potentially biasing listeners to always interpret the ambiguous phonemes in a lexically consistent manner and therefore

generating strong learning effects (Kleinschmidt & Jaeger, 2015; Scharenborg & Janse, 2013). Additionally, listeners always heard the talker's speech in the clear (i.e., absent any background noise), potentially also leading to ceiling levels of learning; an open question is thus whether effects of sentence context might emerge in the context of background noise. Indeed, previous work has shown that LGPL is diminished when listeners encounter simultaneous background noise (Zhang & Samuel, 2014), and the effects of sentence context on online processing of degraded speech are most pronounced at intermediate signal-to-noise ratios (Davis, Ford, Kherif, & Johnsrude, 2011).

However, our results compel us to entertain another possibility, which is that lexically-guided perceptual retuning is relatively immune to the predictability of the critical word. Looking forward, we suggest that additional work is needed to clarify the computational processes that underlie the influences of context on the perception of ambiguous speech. A preceding context may modulate a listener's expectation of how likely an upcoming word is, but it is unclear whether a listener's estimation of the prior probability for a particular phoneme necessarily needs to incorporate all of these cues, particularly when lexical information may provide sufficient disambiguation. Indeed, there are plenty of situations where listeners should not rely too strongly on their expectations about what word is likely to be spoken next. For instance, too strong a reliance on sentence context could lead a listener to inaccurately interpret the innocent question, "Does a bear sit in the woods?" We thus suggest that future work more carefully consider which cues influence listeners' internal estimates of prior probabilities for upcoming phonemes. Furthermore, while preceding context may make some upcoming segment more or less likely to be mapped to a particular phoneme, it is unclear whether this larger prior probability necessarily translates to increased perceptual learning. That is, the degree of phonetic recalibration may not

necessarily be proportional to the activation of a particular phoneme. More detailed computational accounts of the mechanisms underlying perceptual learning will allow more precise hypotheses to be generated and facilitate a better understanding of how the degree of learning interacts with other factors, such as sentence context.

Conclusions

In the current study, we observed robust effects of phonetic recalibration across several experiments that used a non-standard LGPL paradigm. While preceding context influenced how quickly listeners made in-the-moment judgments on target items, it did not affect how much listeners' learned about phonetic idiosyncrasies in these target items. Overall, our results do not provide evidence for a strong influence of preceding context on phonetic recalibration. We suggest that additional work is needed to clarify whether the degree of lexical support influences the process of lexically guided perceptual learning; it will be of particular importance to consider the timing of disambiguating information, the presence of other acoustic information (e.g., noise) in the bottom-up signal, and the factors a listener considers when determining how probable a particular phoneme is *a priori*.

References

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

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

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditoryspeech identification: A McGurk aftereffect. *Psychological Science*, *14*(6), 592–597.

Blank, H., & Davis, M. H. (2016). Prediction errors but not sharpened signals simulate multivoxel fMRI patterns during speech perception. *PLoS Biology*, *14*(11), 1–33.

Boersma, P., & Weenik, D. (2017). Praat: Doing phonetics by computer.

Bonte, M., Correia, J. M., Keetels, M., Vroomen, J., & Formisano, E. (2017). Reading-induced shifts of perceptual speech representations in auditory cortex. *Scientific Reports*, *7*(1), 1–11.

Borsky, S., Tuller, B., & Shapiro, L. P. (1998). "How to milk a coat:" The effects of semantic and acoustic information on phoneme categorization. *The Journal of the Acoustical Society of America*, *103*(5), 2670–2676.

Bowie, C. R., & Harvey, P. D. (2006). Administration and interpretation of the Trail Making Test. *Nature Protocols*, *1*(5), 2277–2281.

Chládková, K., Podlipskỳ, V. J., & Chionidou, A. (2017). Perceptual adaptation of vowels generalizes across the phonology and does not require local context. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(2), 414–427.

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception and Psychophysics*, *70*(4), 604–618.

Colby, S. E., Clayards, M., & Baum, S. R. (2018). The role of lexical status and individual differences for perceptual learning in younger and older adults. *Journal of Speech Language and Hearing Research*, *61*(8), 1855–1874.

Cutler, A., McQueen, J. M., Butterfield, S., & Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boundaries. *Proceedings of Interspeech*, 2056.

Davis, M. H., Ford, M. A., Kherif, F., & Johnsrude, I. S. (2011). Does semantic context benefit speech understanding through "top–down" processes? Evidence from time-resolved sparse fMRI. *Journal of Cognitive Neuroscience*, *23*(12), 3914–3932.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–147.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*(2), 222–241.

de Leeuw, J. R. (2015). jsPsych : A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Drouin, J. R., & Theodore, R. M. (2018). Lexically guided perceptual learning is robust to task-based changes in listening strategy. *Journal of the Acoustical Society of America*, *144*(2), 1089–1099.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing.

Perception & Psychophysics, *67*(2), 224–238.
Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*(4), 1950–1953.
Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125.
Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *14*, 29–56.
Guediche, S., Salvata, C., & Blumstein, S. E. (2013). Temporal cortex reflects effects of sentence context on phonetic processing. *Journal of Cognitive Neuroscience*, *25*(5), 706–718.
Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 460–474.
Jesse, A. (2020). Sentence context guides phonetic retuning to speaker idiosyncrasies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
Jesse, A., & Kaplan, E. (2019). Attentional resources contribute to the perceptual learning of talker idiosyncrasies in audiovisual speech. *Attention, Perception, & Psychophysics*, 1–14.
Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin and Review*, *18*(5), 943–950.
Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3933–3936).
Keetels, M., Schakel, L., Bonte, M., & Vroomen, J. (2016). Phonetic recalibration of speech by text. *Attention, Perception, and Psychophysics*, *78*(3), 938–945.
Kilian-Hütten, N., Valente, G., Vroomen, J., & Formisano, E. (2011). Auditory cortex encodes the perceptual interpretation of ambiguous sound. *Journal of Neuroscience*, *31*(5), 1715–1720.
Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*, *34*(1), 43–68.
Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.
Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54–81.
Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178.
Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review*, *13*(2), 262–268.
Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.
Kraljic, T., & Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition*, *121*(3), 459–465.
Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science*, *19*(4), 332–338.
Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Dartmouth Publishing Group.
Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement : When adults learn new words. *Cognitive Psychology*, *55*, 306–353.

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition*, *174*, 55–70.

Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*(August), 1–16.

Luthra, S., Magnuson, J. S., & Myers, E. B. (2020, June 22). Boosting lexical support does not enhance lexically guided perceptual learning (OSF repository). Retrieved from https://osf.io/eqwja/

Manker, J. (2019). Contextual predictability and phonetic attention. *Journal of Phonetics*, *75*, 94–112.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63.

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.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543–562.

Mazzoni, D., & Dannenberg, R. B. (2015). Audacity.

McAuliffe, M., & Babel, M. (2015). Attention, word position, and perceptual learning. *Proceedings of the 18th International Congress on the Phonetic Sciences*. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0739.pdf

McAuliffe, M., & Babel, M. (2016). Stimulus-directed attention attenuates lexically-guided perceptual learning. *The Journal of the Acoustical Society of America*, *140*(3), 1727–1738.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

McQueen, J. M., Norris, D., & Cutler, A. (2006). The dynamic nature of speech perception. *Language and Speech*, *49*(1), 101–112.

McRae, K., & Boisvert, S. (1998). Automatic semantic similarity priming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*(3), 558–572.

Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PLoS ONE*, *4*(11), 4–8.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46.

Pitt, M. A., & Szostak, C. M. (2012). A lexically biased attentional set compensates for variable speech quality caused by pronunciation variation. *Language and Cognitive Processes*, *27*(7–8), 1225–1239.

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Samuel, A. G. (2016). Lexical representations are malleable for about one second: Evidence for the non-automaticity of perceptual recalibration. *Cognitive Psychology*, *88*, 88–114.

Scharenborg, O., & Janse, E. (2013). Comparing lexically guided perceptual learning in younger and older listeners. *Attention, Perception, & Psychophysics*, *75*(3), 525–536.

Scharenborg, O., Weber, A., & Janse, E. (2015). The role of attentional abilities in lexically guided perceptual learning by older listeners. *Attention, Perception, and Psychophysics*, *77*(2), 493–507.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). afex: Analaysis of Factorial Experiments. R package version 0.21-2. https://CRAN.R-project.org/package=afex.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, *30*(4), 415–433.

van der Zande, P., Jesse, A., & Cutler, A. (2014). Cross-speaker generalisation in two phoneme-level perceptual adaptation processes. *Journal of Phonetics*, *43*, 38–46.

van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1483–1494.

White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents by toddlers. *Developmental Science*, *14*(2), 372–384.

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

Zhang, X., & Samuel, A. G. (2014). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 200–217.

Appendix A

Stimuli. For predictive contexts, cloze ratings for each target word are provided in parentheses.

Target	Predictive Contexts	Neutral Contexts
absent	I had gone to the bathroom when the teacher took roll, which is why he marked me… (0.9)	During the board meeting, the employee advocated that the new campaign tagline be the word…
	Though he was physically present, it was readily apparent that he was mentally… (0.5)	I tuned out because he had babbled on for a while, but I remember he kept using the word…
accent	Even though he has lived here for many years, I can detect a bit of a foreign… (0.75)	After I knocked a jug of water on my paper, the only word I could read was…
	Before he could play the role of a German, the actor needed to learn how to talk with a fake German… (1)	In order to win the game, I would have to get my team to correctly come up with the word…
answer	Even though he did not raise his hand, the teacher called on him for the… (0.75)	You will now hear the target item…
	I told her I didn't want more food, but my mother wouldn't take no for an… (1)	The little girl demanded to know the meaning of the word…
Arkansas	My grandmother lives in Little Rock, which is the capital of… (0.95)	I got annoyed when the typewriter jammed as I was typing the word…
	Before running for president, Bill Clinton was the governor of… (0.75)	I don’t know why, but I am never able to remember the word…
colosseum	He dreamed of being a gladiator and fighting in the… (0.25)	It was on the tip of his tongue, but he could not remember the word…
	The Romans would congregate in a giant amphitheater called the… (0.8)	I've never been able to make out that lyric definitively, but I've always thought the word there was…
currency	In the UK, the pound is used as the local… (0.8)	I had a bizarre dream in which my friends were jumping around a fire and chanting the word…
	If you are traveling abroad, the bank can convert your money into the local… (0.9)	Whoever owned this book before me repeatedly underlined the word…
dinosaur	Though I love the troodon and the pteranadon, the raptor is my favorite kind of… (0.85)	I do not believe he knows the meaning of the word…
	Long before humankind roamed the planet, the world was home to many kinds of… (0.35)	Now that I know Braille, I know that these characters make up the word…
diversity	Troubled that there were no people of color on the faculty, the college talked about ways to promote… (0.8)	The teacher called on me and told me to define the word…
	By referring to America as a "melting pot" of different backgrounds, he hoped to convey the idea that the country values… (0.45)	I only caught the occasional word over the crackle of the PA, but I definitely heard the word…

episode	I love The Walking Dead and eagerly await every new… (0.95)	My ballpoint pen ran out of ink when I was halfway through writing the word…
	I know I need to go to bed, but after that cliffhanger, I have to watch another… (0.75)	I do like that word, but I wonder if it might be better to use the word…
eraser	After copying the math problem incorrectly, he needed to borrow an… (0.85)	The director berated the actor for continually forgetting the word…
	We could not clean the chalkboard after you took the… (0.85)	My one critique of that debater is that he tends to overuse the word…
insane	When the defendant was proven to be mentally ill, he was carted off to a home for the criminally… (0.9)	The five-year-old looked at me blankly when I used the word…
	If a defendant is mentally unable to tell right from wrong, a court might declare them to be legally… (0.75)	A word that came immediately to mind was…
parasite	The literary critic argued that because Dracula feeds off other organisms without conferring any benefit to them, he can be viewed as a… (0.65)	All that was written on the billboard was the word…
	In biology, an organism that leeches off of a different organism is known as a… (0.75)	The only reason I got a bad grade on that German exam was because I couldn't remember the word for…
peninsula	After a week in Morocco, we headed up to the Iberian… (0.7)	Painted on the wall of the modern art museum was the word…
	Florida is not an island but rather a… (0.7)	I was losing for a while, but I took the lead in the board game when I played the word…
pregnancy	The doctor told the future mother not to drink alcohol during her… (1)	Reading over his printed final paper, he was mortified by the highly apparent typo in the word…
	You can find out the gender of the baby halfway through the mother's… (0.8)	When I was a kid, I did not know the meaning of the word…
receipt	After he rang up my coffee order, the employee printed out my… (0.8)	Many of the words had faded over time, but if you look carefully, you can kind of make out the word…
	The boutique will allow you to return anything you bought if you bring it back with the… (0.95)	I really cannot fathom why the only thing on the blackboard is the word…
rehearsal	The band director yelled at the drummer who came late to… (0.2)	There are many great words out there, but I think my all-time favorite word is…
	With opening night on Friday, the director told the actors they would have to work extra hard during… (0.25)	You're mumbling, and the only word I could hear was…
adoption	If you cannot have your own child biologically, there are many children who are available for… (0.9)	He cut random words out of the magazine, finding words like "dandelion" and…
	Before they could legally become the child's guardians, they had to file for… (0.35)	Prominent in the headline on the front page was the word…

brochure	To attract more biology majors, the college included a whole page on the biology program in the annual recruitment… (0.25)	He did not like to admit it, but he did not know the meaning of the word…
	To attract new employees, the recruiter handed out new copies of a tri-fold company… (0.5)	Partway through reading the royal decree, the duke tripped over the word…
definition	In relatively little time, we have gone from watching TV in black and white to being able to watch TV in high… (0.9)	The microphone cut out partway through, but I think the word he was in the middle of was…
	When you use a word that many people won't know, it can be helpful to provide a… (0.45)	The only word I could think of in the moment was…
efficient	To minimize our carbon footprint, we bought bulbs that were highly energy… (0.85)	My writing was too big, and I ran out of room to write the word…
	Because I want to protect the environment, I am looking for a car that is very fuel… (1)	I could only catch the occasional word, but one word I definitely overheard was the word…
friendship	They did not know it when they met at the beginning of camp, but that day marked the beginning of a lifelong… (0.85)	My vision is not great, but I can faintly make out the word…
	There is nothing romantic going on between the two of them; what they have is nothing more than a deep… (0.75)	When I was reading the article, I highlighted the word…
graduation	The valedictorian did not know what to talk about at the junior high… (0.75)	I had trouble remembering the French word for…
	He got a good enough grade on the twelfth grade exit exam that he would be allowed to walk at… (0.75)	I don't want to harp on the point, but I found it really intriguing that the poet used the word…
handshake	The corporate executive greeted me with a firm… (0.9)	I could not believe how many times the writer reused the word…
	I went in for a hug, but in that kind of formal meeting, it might have been more appropriate to go for a… (1)	Hurriedly jotted down on the napkin was the word…
impatient	I am usually accommodating, but after waiting for five hours, even I was feeling… (0.3)	The mother was quite taken aback to learn that her two-year-old daughter already knew the word…
	The car behind me honked the moment the light turned green -- clearly, the driver was feeling rather… (0.5)	The old man wandered the halls, looking at his feet and mumbling the word…
invitation	The bride and groom told their friends to mark their calendars before they mailed a formal wedding… (0.95)	Written prominently in large type at the top of the paper was the word…
	I thought we were good friends, and I was taken aback when I found out he was having a party but I hadn't gotten an… (0.85)	It was unclear if there was any particular reason for him to repeatedly reiterate the word…

ocean	The mighty Amazon river flows into the Atlantic… (1)	The improv comedians wanted a word to riff off of, and one guy in the crowd yelled out the word…
	The majority of the Earth is covered by miles and miles of blue… (0.35)	As if to belabor the point, he kept on repeating the word…
parachute	Before you can jump out of an airplane, you need to have a working… (1)	At long last, the codebreaker figured out that the letters were an anagram for the word…
	The airplane deployed food and equipment to the village, delivering the load by… (0.35)	Written in large print on the album cover was the word…
pediatrician	A doctor for kids is called a… (1)	The only vocabulary word I got wrong was the word…
	An adult needs to go to an adult primary care doctor, but an infant needs to visit a… (0.85)	I don't remember every word of the memo, but it definitely included the word…
permission	Before they were allowed to go on the field trip, the children needed to get a parent to grant them… (1)	The author entertained many options for the title of her book, eventually opting for it to be called…
	Before proposing to his girlfriend of many years, the man went to her father to get… (0.8)	In her paper, the writer contemplated the meaning conveyed by the word…
pressure	When I went in for my appointment, the doctor measured my blood… (0.8)	I could not believe how many times the writer reused the word…
	The mother reminded her daughter not to give in to peer… (1)	I've been working on getting better at calligraphy and am particularly proud of how I wrote the word…
professional	The college athlete trained very hard, hoping one day to be recruited to play as a… (0.55)	The word to evaluate now is the word…
	If you want the job done right, don't go to an amateur; hire a… (0.8)	The director told the actor to be more emphatic, particularly on the word…
vacation	The whole family went to Hawaii for a weeklong… (0.9)	I need help thinking of an antonym for the word…
	After four years without a day off, the couple was ready for a lengthy… (0.8)	Preoccupied by the crying baby, he broke off mid-thought and midway through the word…

Appendix B
Supplemental Analyses

Experiment 1

In Experiment 1, half of the participants (n=40) were recruited through Amazon Mechanical Turk, and half (n=40) participated in person. In a control analysis, we examined whether the extent of perceptual learning differed between settings (online vs. in-lab, coded with a [-1, 1] contrast). Bias and Step were coded as described in the main text, and the maximal random effect structure was identified as the best structure. Results (Table B1) indicate that there were no effects of setting on phonetic categorization.

Effect	**$\chi 2(1)$**	***p***
Bias	43.53	< 0.001
Step	162.96	< 0.001
Setting (online/lab)	1.62	0.20
Bias × Step	3.37	0.07
Bias × Setting	2.12	0.14
Step × Setting	0.09	0.76
Bias × Step × Setting	0.89	0.34

Table B1. Phonetic categorization results from Experiment 1, considering setting as a factor.

To examine whether effects of Bias were stable across blocks, we also ran an analysis that tested for fixed effects of Bias, Step, and Block (first, second; coded with a [-1, 1] contrast). For this analysis, our stepping procedure identified the most parsimonious random effect structure as one with random by-subject intercepts as well as random by-subject slopes for Bias, Bias × Step, and Step × Block. As indicated in Table B2, the effect of Bias did not differ between blocks (i.e., there were no significant interactions with Bias and Block).

Effect	**χ2(1)**	***p***
Bias	53.89	< 0.001
Step	172.78	< 0.001
Block	3.81	0.05
Bias × Step	2.04	0.15
Bias × Block	0.19	0.66
Step × Block	4.84	0.03
Bias × Step × Block	0.47	0.49

Table B2. Phonetic categorization results from Experiment 1, considering block number as a factor.

Experiment 2

In Experiment 2, half of the participants (n=80) were recruited through MTurk, and half participated in an in-lab experiment. As before, we examined whether the extent of learning differed between settings (online vs. in-lab, coded with a [-1, 1] contrast). Bias, Step and Context were coded as described in the main text, and the maximal random effect structure was identified as the best structure. Results indicate that there were no effects of setting on the degree of perceptual learning (i.e., no interactions between Bias and Setting).

Effect	**χ2(1)**	***p***
Bias	18.16	< 0.001
Step	328.34	< 0.001
Context	3.88	0.05
Setting	5.19	0.02
Bias × Step	2.62	0.11
Bias × Context	0.08	0.78
Step × Context	1.28	0.26
Bias × Setting	0.59	0.44
Step × Setting	3.91	0.05
Context × Setting	0.12	0.72
Bias × Step × Context	0.01	0.92
Bias × Step × Setting	1.79	0.18
Bias × Context × Setting	0.38	0.54

Step × Context × Setting	0.06	0.81
Bias × Step × Context × Setting	0.03	0.86

Table B3. Phonetic categorization results from Experiment 2, considering setting as a factor.

To assess whether hearing both biasing conditions (/s/-biased and /ʃ/-biased contexts) may have affected our results, we also conducted an analysis in which we only analyzed data from the first block of phonetic categorization trials, effectively rendering Bias as a between-subjects factor. Random by-subject slopes for and interactions with Bias were removed, but otherwise, the model structure was the same as in the main analysis. Results (Table B4) resembled the results from the main analysis (Table 1 in the main text) – there were main effects of Bias, $p < 0.001$, and Step, $p < 0.001$, but no other significant results.

Effect	**χ2(1)**	***p***
Bias	10.69	< 0.001
Step	315.52	< 0.001
Context	2.43	0.12
Bias × Step	0.05	0.82
Bias × Context	0.01	0.94
Step × Context	0.04	0.83
Bias × Step × Context	1.29	0.26

Table B4. Phonetic categorization results from Experiment 2, considering only the first block of phonetic categorization data (rendering Bias a between-subjects factor).

Finally, we examined whether recalibration was stable across the two blocks. In this analysis, we considered fixed factors of Bias, Step, and Block. (Note that because these were all within-subject manipulations, we lacked the power to simultaneously test for effects of Context in this analysis.) Our model considered random by-subject intercepts as well as random by-subject slopes for Bias, Bias × Step, and Step × Block; our stepping procedure indicated that this was the most parsimonious structure. As shown in Table B5, there were no interactions between Bias and Block.

Effect	**$\chi 2(1)$**	***p***
Bias	23.32	< 0.001
Step	348.90	< 0.001
Block	13.85	< 0.001
Bias × Step	4.06	0.04
Bias × Block	1.95	0.16
Step × Block	1.13	0.29
Bias × Step × Block	0.49	0.48

Table B5. Phonetic categorization results from Experiment 2, considering block number as a factor.

Experiment 3

As before, we conducted an additional analysis that considered only the first block of phonetic categorization data, rendering Bias a between-subjects factor. Results are shown in Table B6. As we found in the main text (Table 2), there were significant effects of Bias and of Step, but no interactions with Context.

Effect	**$\chi 2(1)$**	***p***
Bias	6.47	0.01
Step	274.54	< 0.001
Context	0.83	0.36
Bias × Step	0.28	0.60
Bias × Context	0.18	0.67
Step × Context	0.61	0.43
Bias × Step × Context	0.07	0.80

Table B6. Phonetic categorization results from Experiment 3, considering only the first block of the phonetic categorization data.

We also examined whether phonetic recalibration effects were stable over time by performing an analysis with fixed factors of Bias, Step, and Block. The most parsimonious random effect structure involved random by-subject intercepts as well as random by-subject slopes for Bias, Step, Block, Bias × Step, and Step × Block. As shown in Table B7, we did not observe any interactions between Bias and Block, suggesting comparable retuning effects over time.

Effect	**χ2(1)**	***p***
Bias	33.12	< 0.001
Step	315.05	< 0.001
Block	0.15	0.70
Bias × Step	0.37	0.54
Bias × Block	0.24	0.62
Step × Block	22.17	< 0.001
Bias × Step × Block	0.20	0.66

Table B7. Phonetic categorization results from Experiment 3, considering block number as a factor.

Experiment 4

An additional analysis considered only the first block of phonetic categorization data from Experiment 4. Results are given in Table B8. We observed a Bias × Step × Context interaction, which captures both the fact that phonetic recalibration occurred as well as the slight differences between the size of the learning effect between groups at different parts of the continuum. In particular, participants who read neutral contexts showed slightly larger learning effects on the "sign" end of the continuum, whereas participants who read predictive contexts showed slightly larger learning effects on the "shine" end of the continuum. We also observed a significant effect of Step as well as a Step × Context interaction, the latter which is not theoretically informative.

Effect	**χ2**	***p***
Bias	0.15	0.70
Step	269.19	< 0.001
Context	1.61	0.20
Bias × Step	2.47	0.12
Bias × Context	0.79	0.38
Step × Context	4.34	0.04
Bias × Step × Context	4.34	0.04

Table B8. Phonetic categorization results from Experiment 4, considering only the first block of the phonetic categorization data.

Finally, we examined whether effects of Bias were stable across blocks. This analysis considered fixed factors of Bias, Step, and Block, and the most parsimonious random effect structure had random by-subject intercepts as well as random by-subject slopes for Bias, Bias × Step, and Step × Block. As shown in Table B9, there were no interactions between Bias and Block.

Effect	**χ2(1)**	***p***
Bias	16.98	< 0.001
Step	321.07	< 0.001
Block	1.31	0.25
Bias × Step	6.30	0.01
Bias × Block	2.96	0.09
Step × Block	15.44	< 0.001
Bias × Step × Block	0.54	0.46

Table B9. Phonetic categorization results from Experiment 4, considering block number as a factor.