Hearing is believing: Lexically guided perceptual learning is graded to reflect the quantity of evidence in speech input

Shawn N. Cummings and Rachel M. Theodore

Department of Speech, Language, and Hearing Sciences

University of Connecticut

2 Alethia Drive, Unit 1085

Storrs, CT 06269-1085 USA

Connecticut Institute for the Brain and Cognitive Sciences

337 Mansfield Road, Unit 1272

Storrs, CT 06269-1272 USA

Word count: 15036

Declarations of interest: None

Corresponding author: Rachel M. Theodore (rachel.theodore@uconn.edu)

Abstract

There is wide variability in the acoustic patterns that are produced for a given linguistic message, including variability that is conditioned on *who* is speaking. Listeners solve this lack of invariance problem, at least in part, by dynamically modifying the mapping to speech sounds in response to structured variation in the input. Here we test a primary tenet of the ideal adapter framework of speech adaptation, which posits that perceptual learning reflects the *incremental* updating of cue-sound mappings to incorporate observed evidence with prior beliefs. Our investigation draws on the influential lexically guided perceptual learning paradigm. During an exposure phase, listeners heard a talker who produced fricative energy ambiguous between /ʃ/ and /s/. Lexical context differentially biased interpretation of the ambiguity as either /s/ or /ʃ/, and, across two behavioral experiments (n = 500), we manipulated the quantity of evidence and the consistency of evidence that was provided during exposure. Following exposure, listeners categorized tokens from an *ashi – asi* continuum to assess learning. The ideal adapter framework was formalized through computational simulations, which predicted that learning would be graded to reflect the quantity, but not the consistency, of the exposure input. These predictions were upheld in human listeners; the magnitude of the learning effect monotonically increased given exposure to four, 10, or 20 critical productions, and there was no evidence that learning differed given consistent versus inconsistent exposure. These results (1) provide support for a primary tenet of the ideal adapter framework, (2) establish quantity of evidence as a key determinant of adaptation in human listeners, and (3) provide critical evidence that lexically guided perceptual learning is not a binary outcome. In doing so, the current work provides foundational knowledge to support theoretical advances that consider perceptual learning as a graded outcome that is tightly linked to input statistics in the speech stream.

1. INTRODUCTION

Speech perception is not static; rather, listeners dynamically modify the mapping between acoustic-phonetic cues and speech sound representations to reflect structured regularities in speech input. This flexibility helps listeners solve the lack of invariance problem for speech perception, which arises because many different acoustic patterns are produced for a given speech sound. A primary source of variability in speech input is *who* is speaking. Talker differences in speech production arise from many sources, including physiological (e.g., age, sex), sociophonetic (e.g., dialect, gender, social group), and even idiosyncratic characteristics (Allen et al., 2003; Byrd, 1992; Chodroff & Wilson, 2017, 2018; Fant, 1973; Hillenbrand et al., 1995; Johnson & Beckman, 1997; Klatt, 1986; Munson, 2011; Newman et al., 2001; Peterson & Barney, 1952; Theodore et al., 2009). While not the only source of variability in speech input, talker differences represent a primary contribution towards the lack of invariance between acoustic patterns in the input and an intended linguistic message.

Explicating a theoretical account of how listeners resolve lack of invariance problem is a longstanding challenge in speech perception research (e.g., Liberman et al., 1957, 1967). Many findings implicate adaptation as a necessary component of such a theory (e.g., Heffner et al., 2022; Idemaru & Holt, 2020; Kleinschmidt & Jaeger, 2015; McMurray & Jongman, 2011; Norris et al., 2003; Theodore & Monto, 2019; Tzeng et al., 2021). Rigid, invariant classification models for speech are unable to achieve similar accuracy to human listeners (McMurray & Jongman, 2011), while models that instead allow malleability in perception based on context – such as *who* is speaking – have more success in predicting human responses (Chodroff & Wilson, 2017; Cummings & Theodore, 2022; Kleinschmidt & Jaeger, 2015; Kluender et al., 2019; Luthra et al., 2021; McMurray & Jongman, 2011; Theodore et al., 2019; Theodore & Miller, 2010; Theodore

& Monto, 2019). Indeed, a rich evidence base suggests that listeners constantly adapt to structured variation in speech input (Bradlow & Bent, 2008; Cummings & Theodore, 2022; Drouin et al., 2016, 2016; Giovannone & Theodore, 2021; Idemaru & Holt, 2020; Luthra et al., 2021; Norris et al., 2003; Nygaard et al., 1995; Nygaard & Lunders, 2002; Samuel & Kraljic, 2009; Sidaras et al., 2009; Theodore & Miller, 2010; Theodore & Monto, 2019; Tzeng et al., 2021; Weatherholtz & Jaeger, 2016; Xie et al., 2018). Theoretical accounts of speech perception, including their computational instantiations, stand to benefit from a marriage with behavioral evidence of adaptation in speech perception (Kleinschmidt & Jaeger, 2015).

The ideal adapter framework for speech adaptation (Kleinschmidt & Jaeger, 2015) aims to provide a unifying account of perceptual learning for speech. In this framework, speech sounds are represented as generative distributions of acoustic-phonetic cues formed by long-term experience with the cue-sound mappings of a given language. This framework assumes that talkers' output consists of samples from these distributions, and perception is the result of inferring these distributions given listeners' beliefs of cue-sound mappings. Adaptation is the consequence of updating prior beliefs by integrating observed evidence with existing priors. Computationally, this theory is implemented in a Bayesian belief-updating model. Initial input from a novel talker is processed based on prior knowledge (e.g., knowledge of language- or gender-specific cue distributions). Learning reflects the updating of a category-specific distribution to integrate observed evidence with the prior distribution, weighted by confidence in prior beliefs. The output is posterior distribution beliefs about category means and covariances, reflecting the likelihood of the prior distribution (e.g., formed by global experience) given the observed evidence (e.g., from the specific talker). Iterative updating is predicted to occur until a change in statistics occurs, which may be triggered by a change in context (e.g., a new talker).

Thus, this framework predicts that learning reflects context-dependent (e.g., talker-specific), cumulative integration of listeners' experience with speech input in that context (Kleinschmidt & Jaeger, 2015). This framework has been proposed to explain both distributional learning, reflecting listeners' unsupervised sensitivity to statistical regularities in speech (e.g., Idemaru & Holt, 2014; Liu & Holt, 2015; McMurray et al., 2009; Theodore et al., 2019; Theodore & Monto, 2019), and supervised learning, where lexical context and other disambiguating cues directly guide incorporation of ambiguous phonetic variants into existing speech sound categories (e.g., Bertelson et al., 2003; Drouin & Theodore, 2018; Keetels et al., 2016; Norris et al., 2003; Samuel & Kraljic, 2009; Tzeng et al., 2021)

A core tenet of the ideal adapter framework is that learning is incremental. Iterative integration of new evidence with existing priors yields the prediction of *incremental changes* in category-specific representations. For example, upon encountering a single production of the phoneme /s/ from a novel talker, this framework predicts that perception will be achieved by mapping the production to a category based on expectations of cue-category mappings formed over extensive prior experience with the /s/ category. In a case where the /s/ was atypical (e.g., exhibiting a lower spectral center than would be expected given the novel talker's physiological and social characteristics), this single production may be insufficient to meaningfully shift a listener's category-specific expectations for /s/ because this model gives more weight to globally-derived prior knowledge compared to a single token from a novel talker. And indeed, it may seem anecdotally intuitive for human responses not to rely too heavily on a single point of evidence. However, this framework predicts that repeated observation of the atypical /s/ should *incrementally* shift listeners expectations for this talker in line with increased evidence of an atypical cue-category mapping. This process is predicted to continue until the listener's posterior

beliefs capture the talker's pronunciation well enough that further exposure does not yield further change in expectations. The prediction that the extent of learning should be *graded* to reflect the extent of evidence falls out of this line of reasoning and is formalized in the ideal adapter framework for speech adaptation.

The overarching goal of the present study is to examine whether this tenet of the ideal adapter framework – graded learning as a function of observed evidence – is realized in human behavior. To do so, we use the highly influential lexically guided perceptual learning paradigm, (e.g., Norris et al., 2003; Samuel & Kraljic, 2009) which has recently been linked to predictions made by the ideal adapter framework (Cummings & Theodore, 2022; Liu & Jaeger, 2018, 2019; Luthra et al., 2021; Saltzman & Myers, 2021; Tzeng et al., 2021). In standard form, the lexically guided perceptual learning paradigm requires listeners to complete a lexical decision exposure task followed by a phonetic identification test task. During exposure, listeners hear speech from a single talker wherein canonical sounds are replaced with acoustic energy that is perceptually ambiguous between two speech sounds. A supervisory signal is provided in the form of lexical context because the ambiguous sound is embedded in items that map to a real word referent only if the ambiguous sound is interpreted as one specific category (Ganong, 1980). For example, replacing the canonical /s/ in *personal* with ambiguous spectral energy between /s/ and /ʃ/ allows lexical context to guide listeners to interpret ambiguity as /s/ because *personal* is an English word (but *pershonal* is not). Likewise, replacing the canonical /ʃ/ in *publisher* with the ambiguous spectral energy guides interpretation of the ambiguity as /ʃ/ because *publisher* is a real word (but *publiser* is not). Lexical bias is manipulated between subjects, which allows listeners in each bias group to differentially build expectations for the exposure talker. At test, listeners categorize stimuli drawn from a continuum that spans the categories manipulated during

exposure (e.g., tokens from an *ashi* to *asi* continuum). Evidence of learning manifests as a difference in performance between the two biasing groups at test, indicating that listeners modified the mapping between acoustics and meaning in line with lexical bias during exposure (e.g., more *asi* responses for listeners in the /s/-bias compared to the /ʃ/-bias exposure group).

To date, most research in the lexically guided perceptual learning paradigm has focused on identifying the conditions that are necessary for learning to occur, with learning most often defined as a measurable difference in perception between listeners groups who received differential biasing exposure. Results from this line of investigation have been fruitful in elucidating requisite conditions for learning (Drouin & Theodore, 2018; Norris et al., 2003). For example, listeners must be able to attribute the to-be-learned ambiguity to the speaker, given that learning does not occur when the ambiguity can be attributed to an incidental cause, such as a pen in the talker's mouth (Kraljic et al., 2008; Kraljic & Samuel, 2011; Liu & Jaeger, 2018). In addition, though a lexical decision task is most often used during exposure (e.g., Norris et al., 2003), the learning effect is robust across tasks including passive listening (Jesse, 2021), talker identification (Luthra et al., 2021), syllable and trial counting (McQueen et al., 2006; Samuel, 2016), visual dot monitoring (van Linden & Vroomen, 2007), and amplitude identification (Drouin & Theodore, 2018). Consistent with the goal of identifying necessary conditions for learning, outcomes of learning in this domain are most often considered as a binary result – does any learning occur, or not? – with relatively limited consideration of the *magnitude* of learning that arises from lexically-guided exposure (cf. Cummings & Theodore, 2022; Tzeng et al., 2021).

Visual inspection of the results in the lexically guided perceptual learning literature suggests wide heterogeneity in the magnitude of the learning effect, which potentially reflects numerous methodological differences across the extant literature. One such difference with direct

implications for the ideal adapter framework is the exposure dose – that is, the *quantity* of evidence that listeners receive. In the standard paradigm, exposure dose consists of 20 critical productions (Kraljic et al., 2008; Norris et al., 2003), though may other doses have been used, including two (Liu & Jaeger, 2018), six (Liu & Jaeger, 2018), 10 (Samuel, 2016; Schuhmann, 2012), 11 (Reinisch & Mitterer, 2016), 16 (Liu & Jaeger, 2018; Zheng & Samuel, 2020), 17 (Nelson & Durvasula, 2021), or 40 (Mitterer et al., 2013; Scharenborg & Janse, 2013) critical productions. Though both exposure dose and the magnitude of the learning effect vary widely in the existing literature, few investigations to date have specifically examined the influence of dose on learning, and those that have do not provide optimal tests of the incremental learning predicted by the ideal adapter framework.

For example, Tzeng et al. (2021) examined the relationship between perceptual learning and consistency in exposure input. In their experiment 1, listeners completed the standard lexically guided perceptual learning paradigm and thus received exposure to 20 ambiguous productions in a disambiguating lexical context. In their experiment 2, listeners heard 20 productions of the biased category, but only 10 of the productions contained an ambiguous fricative, with the other 10 consisting of canonical productions. Learning was observed in both experiments; moreover, a significant interaction was observed reflecting a larger learning effect given 20 compared to 10 ambiguous exposures. Though Tzeng et al. (2021) provides compelling evidence in support of graded learning outcomes, their experiment 2 simultaneously manipulated both quantity and consistency of exposure relative to the standard 20 dose condition. Thus, diminished learning could reflect either or both factors, and disambiguation between these hypotheses requires a stricter test of the effects of each in isolation.

Liu and Jaeger (2018) examined learning for exposure doses of two, six, 10, and 16

critical exposures (referred to as causally unambiguous exposures in their experiments 3b, 2b, 1a, and 2a, respectively). The primary analyses tested for a significant learning effect in each dose separately (consistent with the convention to consider learning outcomes as a binary), which showed a statistically significant learning effect for dose conditions of six, 10, and 16 exposures, but not for the dose condition of two exposures. Post-hoc comparisons across the dose conditions revealed no significant difference in learning between the six, 10, and 16 exposure dose conditions even though a positive numeric association between dose and learning was observed, which may reflect insufficient power to detect these differences given that the study was not designed to test this hypothesis specifically.

In an investigation of lexically guided perceptual learning for multiple talkers, Luthra et al. (2021) observed no significant learning effect given exposure to 16 critical productions but did observe learning following exposure to 32 critical productions. Though no significant interaction between learning and exposure dose was observed, Luthra et al. (2021) concluded that 16 exposures is not sufficient to promote learning, consistent with the convention to consider learning as a binary outcome. In an extension of this study, Cummings and Theodore (2022) found evidence of perceptual learning for multiple talkers with exposure doses of both 20 and 40 critical productions, and – as in Luthra et al. (2021) – found no evidence of an interaction between dose and learning. This finding was interpreted as a ceiling effect on learning consistent with 20 exposures providing sufficient evidence to fully accommodate the atypical production.

Though the ideal adapter framework has been invoked as a potential explanatory theory for numerous findings in the speech perception literature (e.g., Liu & Jaeger, 2019; Luthra et al., 2021; Saltzman & Myers, 2021; Theodore et al., 2019; Theodore & Monto, 2019; Tzeng et al., 2021), it is often invoked in general terms – what might be considered a "verbal theory" (van

Rooij & Blokpoel, 2020) – instead of drawing on its formal computational architecture. As outlined in Kleinschmidt and Jaeger (2015), what this framework predicts is neither generic nor absolute – instead, specific predictions from this framework are contingent on numerous assumptions (e.g., how listeners represent speech sounds, prior knowledge of speech sound representations, how confident the system is in its prior knowledge) and specific aspects of the to-be-explained behavior (e.g., the degree to which input deviates from prior expectations, the amount of evidence a listener receives). To preview some of the results from computational simulations for the current investigation (reported below), even an iterative learning algorithm that reflects a cumulative integration of observed evidence can yield predictions that would suggest a binary account of adaptation (e.g., learning requires some critical number of exposures) if, for example, confidence in prior knowledge is very high. Likewise, an iterative learning algorithm can also yield the prediction that *no* behavioral change will be observed if, for example, the novel input is perfectly in line with prior expectations. Though a "verbal theory" level engagement with this theory is not without merit, it fails to capitalize on the computational instantiation of the ideal observer framework, which can provide specific, fine-grained predictions for human behavior. Additionally – and critically – invocation at the level of a verbal theory often fails to clarify researcher assumptions that are critical for linking observed behavior to the ideal adapter framework.

In this context, the goal of the current work is two-fold. First, we aimed to test a primary tenet of the ideal adapter framework, which states that adaptation is incremental to reflect the *quantity* of evidence in the input (Kleinschmidt & Jaeger, 2015). Second, we aimed to test whether adaptation also reflects *consistency* of evidence in the input (Tzeng et al., 2021). To do so, we first present a series of computational simulations using the ideal adapter framework in

order to be explicit in the assumptions guiding our test of this framework and to generate fine-grained predictions for human behavior (Kleinschmidt, 2017; Kleinschmidt & Jaeger, 2015, 2016; Theodore & Monto, 2019). We then test whether the predicted influence of exposure quantity (experiment 1) and exposure consistency (experiment 2) on adaptation is observed in human listeners.

2. MODEL SIMULATIONS

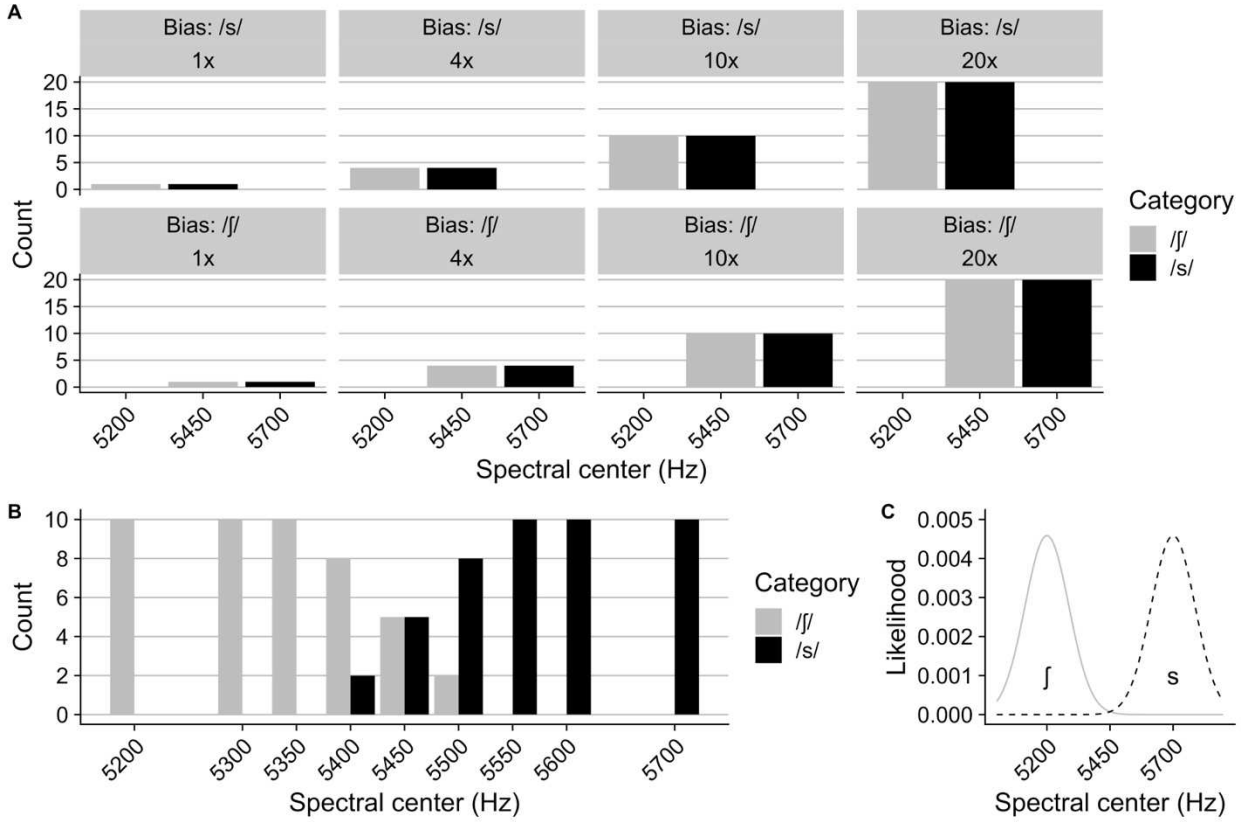
Simulations were implemented using the beliefupdatr (Kleinschmidt, 2017) and slopeExtractR (Monto, 2018) packages in R. All code to reproduce the simulations reported here is available on the Open Science Framework (OSF) repository for this manuscript: https://osf.io/zkbng/. In this model, prior specifications consist of the mean and variance of two categories (/ʃ/ and /s/) along with a confidence parameter that estimates the number of direct observations of the prior specification. The model input is trial-by-trial observations of an acoustic-phonetic parameter (e.g., spectral center) and the response category (e.g., /s/ or /ʃ/). Using this input, the learning algorithm (explicated fully in Kleinschmidt and Jaeger, 2015) updates the category-specific distributions on each trial by integrating the observed evidence (i.e., the spectral center and response) with the prior distribution, weighted by confidence. For each trial, the output is the posterior distribution given the observed evidence. The algorithm is iterative at each trial and thus simulates how beliefs change over time given new evidence. Using this model, we performed two sets of simulations, one that examined adaptation as a function of the *quantity* of the exposure and one that examined adaptation as a function of the *consistency* of the exposure. The only difference between the quantity and consistency simulation procedures is that the consistency simulations presented the system with inconsistent evidence of a talker's phonetic implementation of the biased phonetic category whereas the quantity simulations

provided consistent evidence of "ambiguous" production of the biased phonetic category. Each is described in turn, below.

2.1. Quantity simulations

Our procedure (1) simulated input in the standard lexically guided perceptual learning paradigm in which listeners hear clear variants of one category and ambiguous variants of a different category and (2) simulated a different quantity of evidence across four dose conditions. We implemented this procedure as follows. First, prior specifications were set to reflect a "typical talker" based on acoustic data for the /ʃ/-/s/ contrast provided in Newman et al. (2001). Specifically, mean spectral center was set to 5200 Hz for /ʃ/ (*SD* = 87 Hz) and 5700 Hz for /s/ (*SD* = 87 Hz), as shown in Figure 1, panel C. Second, we generated 400 lists (each simulating a

Figure 1. Input for the quantity simulations. Panel A shows the exposure input for each bias by dose condition. Panel B shows the test input, which was constant across all bias by dose conditions. Panel C shows the prior specifications for all simulations.



unique listener) specifying trial-level spectral center and corresponding response category, reflecting 100 lists for each of four dose conditions, reflecting one (1x), four (4x), 10 (10x), or 20 (20x) critical exposures to an atypical production. In each of six experiments, Cummings and Theodore (2022) found no evidence that learning differed given exposure to 20 vs. 40 critical exposures, consistent with 20 exposures being sufficient to achieve a ceiling learning effect. Accordingly, the doses examined in the current work sample doses where the quantity of exposure is hypothesized to yield a measurable effect on learning outcomes. All lists consisted of exposure input followed by test input. Within each dose condition, 50 lists simulated /s/-bias exposure and 50 lists simulated /ʃ/-bias exposure. The only difference among the 50 lists within each bias by dose condition was the order in which trial-level observations were presented in each phase; each list consisted of a separate random order of the exposure input followed by a separate random order of the test input to simulate 50 unique subjects in each cell. As shown in Figure 1, panel A, exposure input in each list consisted of equal numbers of /s/ and /ʃ/ observations for a given dose. For example, the 1x dose lists contained two exposure trials (one /s/ and one /ʃ/ observation) and the 4x dose lists contained eight exposure trials (four /s/ and four /ʃ/ observations). For all dose conditions, trial-level input specified the acoustic-phonetic parameter for the clear category to match the prior mean of the respective category (5200 Hz for /s/, 5700 Hz for /s/) and the acoustic-phonetic parameter for the ambiguous category to the midpoint frequency between the two categories (5450 Hz). Figure 1, panel A also shows that the exposure input simulated perfect acceptance of the ambiguous input as the intended category. To make this procedure more concrete, consider the simulated exposure input for the 4x dose, /s/-bias condition. Trial-level observations contained four observations of 5200 Hz labeled as /ʃ/, reflecting four observations of the /ʃ/ category that were perfectly in line with prior expectations.

Trial-level observations also contained four observations of 5450 Hz labeled as /s/, which simulates four observations in which lexical context led to a successful map between the midpoint spectral center and the /s/ category. Now consider the simulated exposure input for the 4x dose, /ʃ/-bias condition. The trial-level observations here included four observations of 5700 Hz labeled as /s/ and four observations of 5450 Hz labeled as /ʃ/. Accordingly, the evidence simulated four observations in which the /s/ category perfectly aligned with prior expectations and four observations in which lexical context led to a successful map between the midpoint spectral center and the /ʃ/ category.

Following the exposure input, each list contained 90 observations that simulated test input. As shown in Figure 1, panel B, these 90 observations were constant across all dose and bias conditions and reflected 10 observations of nine test steps. Spectral center of the test steps spanned the range of spectral center frequencies in the prior specifications and was set to simulate the sampling across the /ʃ/-/s/ space of the test continuum that was used in the behavioral experiments. Responses for the test input simulated a typical psychometric response function. Specifically, responses to steps 1 – 3 and steps 7 – 9 reflected perfect categorization of endpoints, responses to the step 5 reflected ambiguous categorization of the continuum midpoint, and responses to steps 4 and 6 were intermediate between the midpoint and the corresponding endpoint.

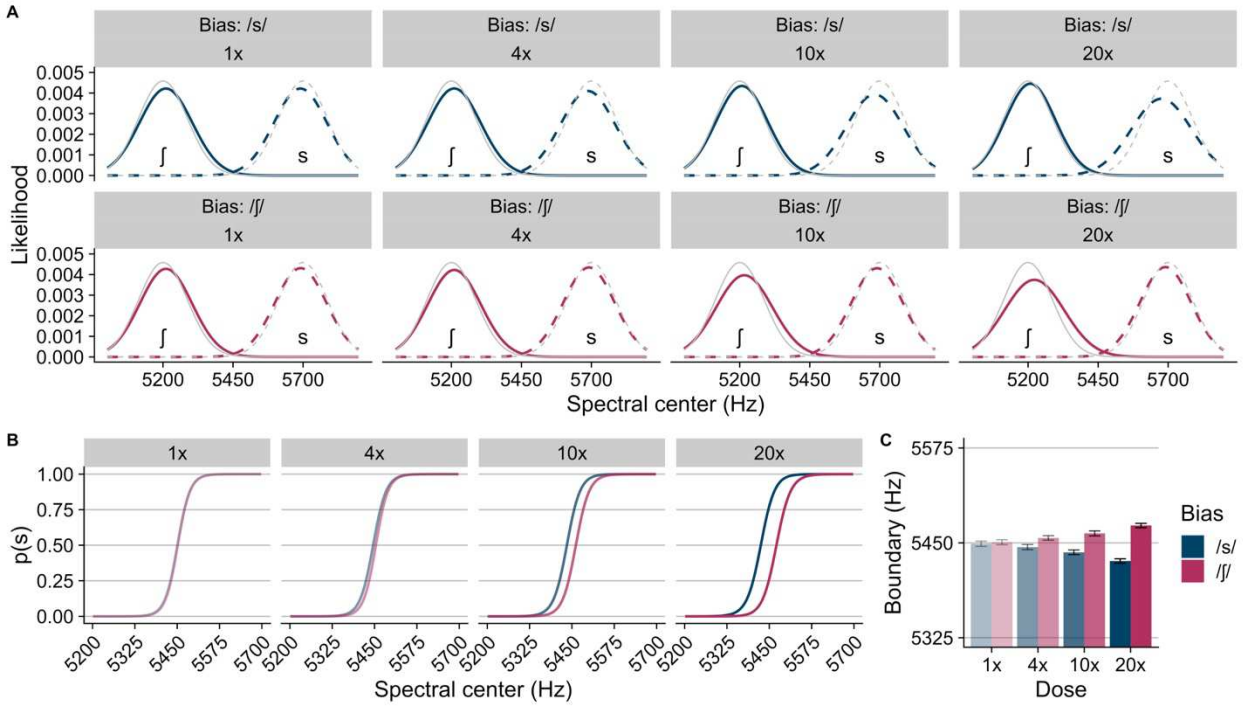
Third, three simulations were performed for these lists, representing three unique confidence specifications. In this model, the confidence parameter reflects a pseudo-count of the number of observations of the prior specification. The confidence parameter interacts with the learning algorithm in that the degree of belief-updating (i.e., the degree to which the priors change given new evidence) is influenced by confidence in the priors; lower confidence values

yield a system that is more flexible (and thus exhibits relatively *strong* belief-updating given new evidence) while higher confidence values yield a system that is more stable (and thus exhibits relatively *weak* belief-updating given new evidence). The three simulations reflected confidence values of 50, 150, and 300, which sample the range of inferred confidence reported in previous work. Finally, for each simulation and each list, we (1) extracted the posterior distribution (i.e., the updated beliefs) at test trials 9, 45, and 90, (2) calculated the predicted categorization function for each of these test trials based on the extracted posterior distribution, and (3) extracted the category boundary of each categorization function, defined as the spectral center corresponding to 0.5 proportion /s/ responses. The three selected test trials reflect simulated performance at the end of the first test cycle (i.e., after one presentation of the nine simulated continuum steps), performance at the end of the test phase used in the behavioral experiments reported below (i.e., after five cycles of the nine simulated continuum steps), and performance at the end of a secondary test phase (i.e., after an additional five cycles of the nine continuum steps). Three different trials were selected because belief-updating in this model is iterative throughout the entirety of the simulation; that is, the model continues to update beliefs even throughout the simulated test phase. As we revisit in the discussion, there is a growing body of evidence indicating that the magnitude of the lexically guided perceptual learning effect attenuates throughout the test phase (Giovannone & Theodore, 2021; Liu & Jaeger, 2018; Tzeng et al., 2021), consistent with adaptation occurring given exposure to the test stimuli themselves.

The results of the quantity simulations are shown in Figure 2. To streamline the exposition, here we present the results of the simulations performed with the confidence level of 300 (i.e., the highest confidence in the priors and thus the most stable/least flexible system) at test trial 45 (i.e., at the end of test phase used in the behavioral experiments of the current work).

The results of the simulations for the other test trials and with the other confidence levels are presented in full in the Supplementary Material; the qualitative patterns presented here held across all confidence specifications and examined trials. Figure 2, panel A shows the extracted beliefs (i.e., the posterior distribution) for one simulated subject in each bias by dose condition. The prior specification at the start of the simulation is shown in gray. Consider first the beliefs for the 20x dose, /s/-bias condition. Compared to the prior specification, the updated belief of the /s/ category has shifted to have a slightly lower mean spectral center and, more notably, a wider standard deviation. This pattern is consistent with beliefs changing to reflect evidence that an "ambiguous" spectral center value is produced for the /s/ category. In contrast, the updated beliefs for the /ʃ/ category show minimal change from the prior beliefs, reflecting evidence that aligned with prior beliefs. Now consider the beliefs for the 20x, /ʃ/-bias condition. Here the

Figure 2. Results of the quantity simulations. Panel A shows the updated beliefs for one simulated subject in each bias by dose condition in addition to the prior beliefs (in gray). Panel B shows the inferred identification function for each set of updated beliefs. Panel C shows the mean category boundary as derived from the inferred identification functions across the 50 simulated subjects in each bias by dose condition; error bars indicate standard error of the mean.



updated belief of the /ʃ/ category deviates from prior expectations, with the updated belief reflecting a higher mean spectral center and increased variability compared to the prior belief. No substantial change to the /s/ belief is observed, consistent with receiving evidence that confirmed the prior belief.

Figure 2, panel B shows the optimal identification function given the updated beliefs for the simulated subjects (panel A). There is no displacement between the /s/-bias and /ʃ/-bias functions for the 1x dose condition, consistent with shared beliefs for the /s/ and /ʃ/ categories. In contrast, displacement is observed for the 4x, 10x, and 20x dose conditions such that the /s/-bias function is shifted towards a lower spectral center compared to the /ʃ/-bias function. This pattern mimics the key learning effect in lexically guided perceptual learning and demonstrates that this effect can be modeled as the consequence of integrating observed evidence with prior beliefs. Moreover, the displacement between the /s/- and /ʃ/-bias functions monotonically increases across dose conditions, yielding the prediction that the magnitude of the learning effect will be positively associated with the quantity of evidence provided during exposure.

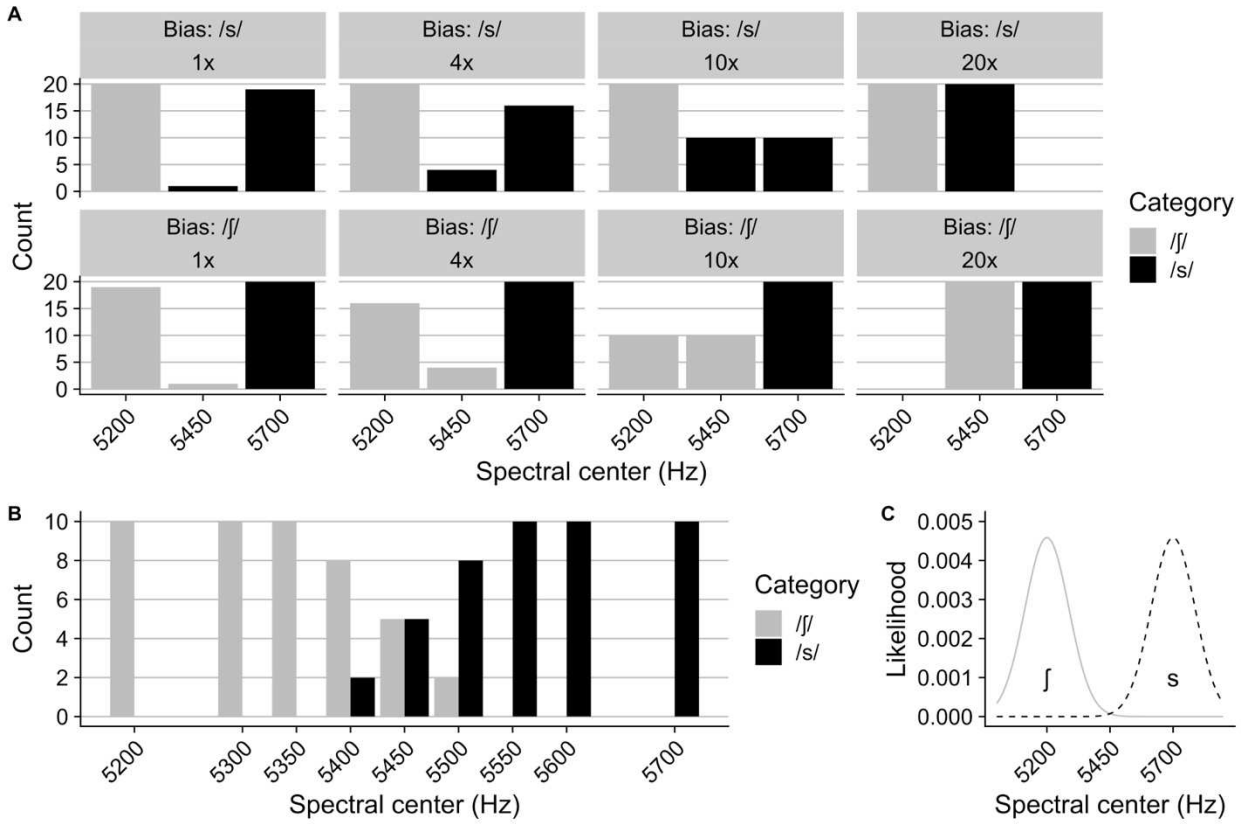
Finally, Figure 2, panel C shows the mean category boundary across all 50 simulated subjects in each dose by bias condition. The graded learning effect in line with graded evidence (i.e., dose) holds not only for the individual subject simulations presented in Figure 2, panels A and B, but also when considering the predicted category boundaries across all 50 simulated listeners in each bias by dose condition. Specifically, predicted category boundaries for the /s/-bias simulations are at a lower center frequency compared to the /ʃ/-bias simulations, and the difference in the boundary between the two bias conditions grows as dose increases.

2.2. Consistency simulations

The consistency simulation procedure was identical to the quantity simulation procedure

with one key exception; namely, the simulated exposure input presented 20 observations of each category. As shown in Figure 3, panel A, this yielded exposure evidence that preserved the dose manipulation of the quantity simulations while also presenting the system with inconsistent evidence of a talker's phonetic implementation of the biased phonetic category. As an example, consider the input for the 10x dose, /s/-bias condition. The simulated input contained 20 observations of the 5200 Hz spectral center labeled as /ʃ/, thus simulating 20 observations of this category that were perfectly aligned with prior expectations. The input also contained 20 observations that were labeled as /s/; 10 observations had a spectral center in line with prior expectations (5700 Hz) and 10 observations had an ambiguous spectral center based on prior expectations (5450 Hz). Accordingly, the quantity of evidence in support of incorporating the

Figure 3. Input for the consistency simulations. Panel A shows the exposure input for each bias by dose condition. Panel B shows the test input, which was constant across all bias by dose conditions. Panel C shows the prior specifications for all simulations.



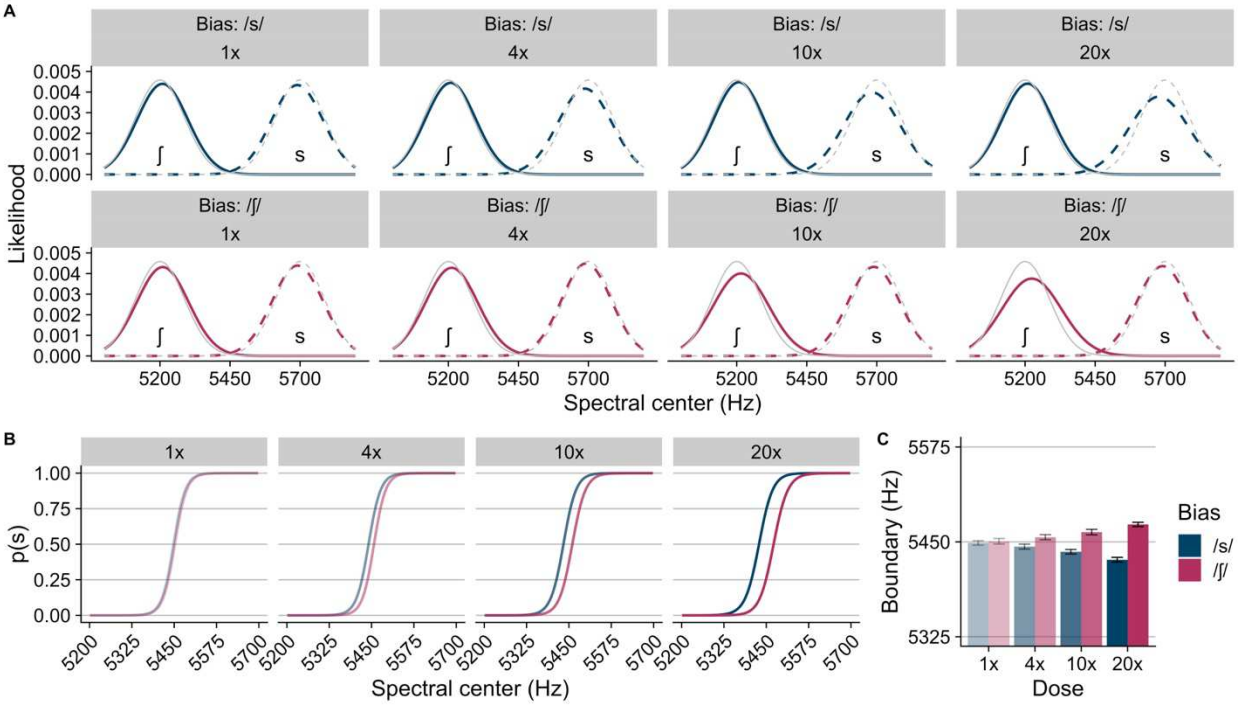
ambiguous spectral center into the /s/ category matched the 10x dose condition in the quantity simulations, but the global evidence in support of this shift was inconsistent. In the 1x dose condition, the input simulated one observation of the ambiguous spectral center along with 19 observations of the same category that matched the category prior. In the 4x dose condition, the input simulated four observations of the ambiguous spectral center along with 16 observations of the same category that matched the category prior. The 20x dose condition was included for completeness; we note that the input for these simulations was identical to that of the parallel quantity simulations.

The results of the consistency simulations are shown in Figure 4. These results reflect simulations performed with the confidence level of 300 at test trial 45. Full results are shown in the Supplementary Material; the qualitative patterns presented here converged across all simulated confidence levels and test trials. We imagine that the reader will note a striking similarity in the results of the consistency simulations (Figure 4) compared to the quantity simulations (Figure 2). Indeed, the predicted identification functions (Figure 4, panel B) and category boundaries (Figure 4, panel C) are strikingly similar both qualitatively *and* quantitatively to those observed for the quantity simulations. Specifically, the magnitude of the learning effect (i.e., the displacement between the simulated /s/- and /ʃ/-bias conditions) is graded in line with the quantity of evidence and does not appear to be diminished by the lack of consistency in the simulated input.

Why these predicted patterns emerge can be understood by considering the updated beliefs shown in Figure 4, panel A. For the quantity simulations, updated beliefs for a given category only differ from the prior belief when the simulated evidence was inconsistent with the prior; that is, the updated beliefs for the /s/ category differ from the prior in the /s/-bias

simulations and the updated beliefs for the /ʃ/ category differ from the prior in the /ʃ/-bias simulations. In each case, the updated beliefs differ not only by a shift in the mean but also a widening of the standard deviation. Recall that in this computational framework, receiving evidence that aligns with expectations yields no change to beliefs. This is robustly apparent in the model beliefs for the category that reinforced the prior (i.e., the /ʃ/ category in the /s/-bias simulations and the /s/ category in the /ʃ/-bias simulations), and these tenets explain why the model predictions converge between the quantity and consistency simulations.

Figure 4. Results of the consistency simulations. Panel A shows the updated beliefs for one simulated subject in each bias by dose condition in addition to the prior beliefs (in gray). Panel B shows the inferred identification function for each set of updated beliefs. Panel C shows the mean category boundary as derived from the inferred identification functions across the 50 simulated subjects in each bias by dose condition; error bars indicate standard error of the mean.



To make this explicit, consider the 10x dose, /s/-bias condition. Receiving evidence of 10 "ambiguous" productions of the /s/ category required the model to update beliefs about the /s/ category, including a widening of the expected variance of this category. The inclusion of 10 "clear" productions of the same category is already accommodated by the belief that the category

has a large standard deviation. Accordingly, additional interspersed evidence of 10 "clear" productions of /s/ does not require any *differential* belief-updating compared to widening category expectations given "ambiguous" exposure alone.

Viewed collectively, the results of the quantity and consistency simulations make two clear predictions for human behavior. First, both sets of simulations predict that the magnitude of the lexically guided perceptual learning effect will be graded in response to the exposure dose, with learning increasing as evidence increases. Second, comparing across the quantity and consistency simulations for a given exposure dose leads to the prediction that learning will not be affected by exposure consistency. In the experiments presented below, we test these two predictions in human listeners.

3. EXPERIMENT 1

The results of the computational simulations confirm a key tenet of the ideal adapter model of adaptation; namely, that degree of adaptation (i.e., learning) is graded to reflect quantity of input evidence. The goal of experiment 1 is to test this prediction with human listeners. Eight groups of listeners completed a lexical decision exposure phase followed by a test phase. Across listener groups, we parametrically manipulated lexical bias (/s/-bias vs. /ʃ/-bias) and exposure dose (1 vs. 4 vs. 10 vs. 20 critical productions). Within each dose, the learning effect was measured as the difference between the two bias conditions at test. If lexically guided perceptual learning is graded to reflect quantity of evidence in the input, as predicted by the ideal adapter framework, then the magnitude of the learning effect will monotonically scale with exposure dose.

3.1. Methods

3.1.1. Participants

The sample size and inclusion/exclusion criteria were preregistered; the preregistration is available on the OSF repository for this manuscript (https://osf.io/zkbng/).[1] Participants (n = 400) were recruited from the Prolific participant pool according to the following criteria: between 18 – 35 years of age, born in and currently residing in the United States, monolingual English speaker, no history of language related disorders, and a Prolific approval score ≥ 98 based on completion of ≥ 10 studies. Moreover, no participant took part in any previous lexically guided perceptual learning study in our laboratory. Participants were randomly assigned to one of eight cells formed by crossing lexical bias (/s/-bias vs. /ʃ/-bias) and exposure dose (1x vs. 4x vs. 10x vs. 20x) yielding 50 participants in each cell.

An additional 13 participants were tested but excluded from analyses due to failure to pass the headphone screen (n = 6), failure to achieve ≥ 80% accuracy during the lexical decision exposure phase (n = 6), or failure to respond to ≥ 10% of the trials (n = 1). The final sample included 172 men, 226 women, and two individuals who declined to report gender. The mean age of participants was 26 years (SD = 5 years, *range* = 18 – 35 years). As described in the

[1] The experiments presented here deviated from the preregistered protocol in two ways. First, the inclusion criterion for lexical decision accuracy during exposure was set to ≥ 80% correct instead of ≥ 90% correct. This is because we inaccurately set the inclusion criterion in the preregistration; it was adjusted to ≥ 80% to be more consistent with the convention for lexical decision accuracy. The results presented in the main text hold when limited to those who met the preregistered inclusion criterion, as can be viewed by executing the script provided on the OSF repository for this manuscript. Second, we only tested listeners in the 10x dose condition in experiment 2 instead of also testing listeners in the 1x and 4x dose conditions. We made this decision for two reasons. First, the results of experiment 1 provided no strong evidence of learning for the 1x and 4x dose conditions. Though a statistically significant learning effect was observed for the 4x dose condition, the magnitude of this effect was small, and there was no significant interaction between bias and dose for the 1x vs. 4x contrast. Second, testing the consistency hypothesis in experiment 2 requires testing whether the magnitude of learning *when it occurs* is influenced by consistency of evidence for a given dose. Because there was no strong evidence of learning for the (consistent) 1x and 4x dose conditions in experiment 1, including these conditions in experiment 2 did not seem a justifiable use of lab resources; accordingly, only the 10x dose condition was included in experiment 2.

Supplementary Material, a priori power analyses indicated that this sample size, reflecting 50 participants in each between-subjects condition, provided high power to detect our effect sizes of interest.

3.1.2. Stimuli

The stimuli were identical to those used in Tzeng et al. (2021), to which the reader is referred for comprehensive details on stimulus construction. In brief, the stimulus set consisted of 240 exposure tokens [(20 /s/ words x 2 variants) + (20 /s/ words x 2 variants) + 60 filler words + 100 nonwords] and nine test tokens. Exposure tokens included auditory recordings of 100 English words, 20 containing a single instance of /s/ and no occurrence of /ʃ/ (e.g., *rehearsal*), 20 containing a single /ʃ/ and no occurrence of /s/ (e.g., *publisher*), and 60 that contained no instances of either /s/ or /ʃ/ (e.g., *ballerina*). Two variants of the /s/ and /ʃ/ words were created, one that contained the natural production of /s/ or /ʃ/ (the clear variant) and one in which the natural production of /s/ or /ʃ/ was replaced with a digital mixture of a natural /s/ and /ʃ/ production that was judged to be perceptually ambiguous between /s/ and /ʃ/ (the ambiguous variant). As described in Tzeng et al. (2021), the ambiguous variant consisted of a custom digital mixture for each /s/ and /ʃ/ word. Exposure tokens also included auditory recordings of 100 nonwords that contained no instances of either /s/ or /ʃ/ (e.g., *baliber*).

Test tokens consisted of a nine-step continuum that perceptually ranged from /ɑʃi/ to /ɑsi/. The fricative portion of the test continuum was created by digitally mixing energy from natural /ʃ/ and /s/ productions in different weights to yield continuum steps that ranged from 80% /ʃ/ - 20% /s/ to 20% /ʃ/ - 80% /s/ in seven equidistant units for the midpoint steps (steps 2 – 8). The endpoint steps reflected a 100% /ʃ/ - 0% /s/ mixture (step 1) and a 0% /ʃ/ - 100% /s/ mixture (step 9). All exposure and test tokens were produced by a single female talker, referred to as “f1”

in Tzeng et al. (2021).

3.1.3. Procedure

All experiments presented here were web-based studies hosted on the Gorilla platform (Anwyl-Irvine et al., 2020). After providing informed consent, participants completed a headphone screen, an exposure phase, and a test phase. The headphone screen used tasks reported in Woods et al. (2017) and Milne et al. (2021), which are brief, dichotic listening tasks developed to screen for headphone use in web-based experiments.

The exposure phase consisted of a lexical decision task. All participants heard 60 filler words and 100 nonwords during exposure. In addition, listeners in the /s/-bias conditions heard ambiguous variants of /s/ words and clear variants of /ʃ/ words, and listeners in the /ʃ/-bias conditions heard ambiguous variants of /ʃ/ words and clear variants of /s/ words. As shown in Table 1, we manipulated the number of /s/ and /ʃ/ words across dose conditions such that (1) equal numbers of the critical /s/ and /ʃ/ words were presented for a given dose condition and (2) the number of critical words differed across dose conditions. For each participant, items were randomly sampled from the full stimulus set. For example, each listener in the /s/-bias, 1x dose condition heard one ambiguous /s/ word that was randomly sampled from the full set of 20 ambiguous /s/ words and one clear /ʃ/ word that was randomly sampled from the full set of 20 clear /s/ items. Listeners in the 20x dose conditions thus heard the full set of items appropriate for their bias condition. On each trial, listeners heard one item and were asked to indicate whether the item was a real English word or not by clicking on one of two buttons labeled either "Yes" or "No." Assignment of button labels was counterbalanced across listeners within each cell. No feedback was provided, and trials were separated by 1000 ms, timed from the participant's response to the onset of the next auditory stimulus.

Table 1. Distribution of exposure trials by item type and total number of trials for each bias and dose condition in each experiment.

Experiment	Bias	Dose	Words				Filler	Nonwords	Trials
			/s/		/ʃ/				
			Ambiguous	Clear	Ambiguous	Clear			
1	/s/	1x	1	-	-	1	60	100	162
		4x	4	-	-	4	60	100	168
		10x	10	-	-	10	60	100	180
		20x	20	-	-	20	60	100	200
	/ʃ/	1x	-	1	1	-	60	100	162
		4x	-	4	4	-	60	100	168
		10x	-	10	10	-	60	100	180
		20x	-	20	20	-	60	100	200
2	/s/	10x	10	10	-	20	60	100	200
	/ʃ/	10x	-	20	10	10	60	100	200

Following the exposure phase, all listeners completed a test phase that was identical across all dose and bias conditions. The primary test phase consisted of 45 trials of a phonetic identification task, reflecting five cycles of the nine steps of the test continuum. For each participant, each cycle was a separate randomized order of the nine continuum steps. Following the primary test phase, participants completed an additional five test cycles that were used in exploratory analyses (presented in the Supplementary Material) to examine the influence of dose on learning over time, given past research that has shown that the learning effect in the standard dose condition atrophies over longer test sessions (e.g., Giovannone & Theodore, 2021; Liu & Jaeger, 2018; Tzeng et al., 2021). As for the primary test phase, each cycle in the secondary test phase was a separate randomized order of the nine continuum steps for each participant.

On each trial, listeners heard one test token and were asked to indicate its identity as quickly and as accurately as possible by clicking on one of two buttons labeled either “ashi” or “asi.” Assignment of button labels was counterbalanced across listeners. As with the exposure

phase, no feedback was provided at test. Trials were separated by 1000 ms timed from the participant's response to the onset of the next auditory stimulus. The entire procedure lasted approximately 15 minutes and participants were paid $2.50 for their participation.

3.2. Results

Trial-level data and analysis code for all experiments reported here are available on the OSF repository for this manuscript: https://osf.io/zkbng/. Performance during the exposure phase was analyzed in terms of percent correct lexical decision accuracy. An item was considered correct if the response matched the intended item type (i.e., a response was correct if the participant responded "word" to the /s/ word, /ʃ/ word, or filler word items or if the participant responded "nonword" to the nonword items). As shown in Table 2, mean lexical decision was near ceiling for each bias by dose condition, as expected given that high accuracy during exposure was an inclusion criterion for participation.[2]

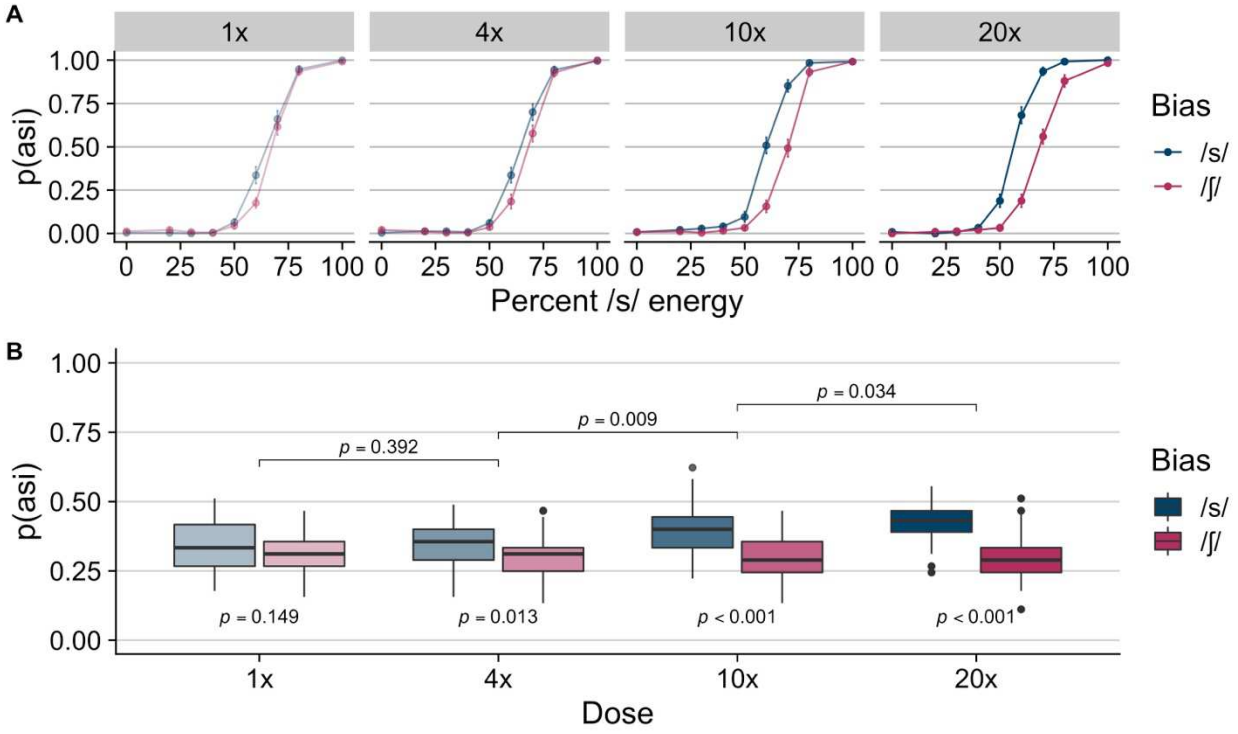
Performance during test was analyzed in terms of *asi* responses for the phonetic identification task. To visualize performance, Figure 5, panel A shows mean proportion *asi*

[2] In the Supplementary Material, we present the results of a secondary analysis of performance during the exposure phase that examined lexical decision accuracy (percent correct) as a function of dose and bias for each of the four item types presented during exposure (i.e., /s/ words, /ʃ/ words, filler words, and nonwords). In brief, the results of this analysis showed no significant effect of dose on lexical decision accuracy nor any interaction between dose and bias for any of the four item types. The lack of an influence of dose on lexical decision accuracy mitigates the concern that the slightly unequal number of word and nonwords items in the 1x, 4x, and 10x dose conditions (see Table 1) may have introduced a response bias that affected lexical decision accuracy. Accuracy for /s/ words was slightly lower for listeners in the /s/-bias conditions who heard ambiguous variants (*mean* = 95.5, *SD* = 14.5) compared to listeners in the /ʃ/-bias conditions who heard clear variants (*mean* = 99.2, *SD* = 3.4), suggesting a slight degree of reluctance in endorsing ambiguous variants as lexical items. Similarly, accuracy for /ʃ/ words was slightly lower for listeners in the /ʃ/-bias conditions (*mean* = 97.9, *SD* = 11.2) compared to the /s/-bias conditions (*mean* = 99.8, *SD* = 1.2). Though slight differences in accuracy were observed for the ambiguous and clear variants of the critical /s/ and /ʃ/ words, accuracy in all cases was near ceiling, indicating that the ambiguous variants were overwhelming endorsed as lexical items. The reader is referred to the Supplementary Material for a comprehensive reporting of this secondary analysis.

Table 2. Mean lexical decision accuracy (percent correct) for each dose and bias condition in each experiment. Means reflect grand means calculated over by-subject means. Standard deviations are provided in parentheses.

		Dose			
Experiment	Bias	1x	4x	10x	20x
1	/s/	96.3 (3.7)	96.3 (3.3)	96.8 (3.3)	97.6 (2.1)
	/ʃ/	97.1 (2.5)	96.6 (2.8)	96.8 (2.6)	97.1 (3.1)
2	/s/	-	-	96.6 (2.7)	-
	/ʃ/	-	-	97.5 (1.9)	-

Figure 5. Results of experiment 1. Panel A shows mean proportion *asi* responses at test for each continuum step separately for each bias by dose condition. Means reflect grand means calculated over by-subject averages; error bars indicate standard error of the mean. Panel B shows the boxplot distribution of mean proportion *asi* responses across listeners in each bias by dose condition collapsing across continuum step. Annotations below the boxplot distributions indicate *p*-values associated with the main effect of bias in each dose condition. Annotations above the boxplot distributions indicate *p*-values associated with the bias by dose interaction, as described in the main text.



responses for each continuum step separately for each bias by dose condition; means reflect grand means calculated over by-subject means. Figure 5, panel B shows the distribution of by-

subject means for each bias by dose condition collapsing across continuum step. Inspection of Figure 5 suggests that the magnitude of the learning effect (i.e., the difference between the two bias conditions) monotonically scales with exposure dose from a limited (or perhaps absent) learning effect in the 1x dose condition to a relatively maximal learning effect in the 20x dose condition.

To examine these patterns statistically, trial-level responses (0 = *ashi*, 1 = *asi*) were submitted to a series of generalized linear mixed effects models (GLMMs) with the binomial response family as implemented in lme4 (Bates et al., 2015); the Satterthwaite approximation of degrees of freedom was used to evaluate statistical significance using lmerTest (Kuznetsova et al., 2017).[3] The first set of models tested for an effect of bias in each dose condition separately. Accordingly, these models reflect the standard approach to identifying learning in this paradigm; that is, to identify conditions under which any effect of bias occurs. For each of these models, the fixed effect was bias and the random effects structure consisted of random intercepts by subject and random slopes by subject for continuum step. Bias was entered into the model as a mean-centered contrast (/ʃ/ = -0.5, /s/ = 0.5); continuum step was entered into the model in terms of percent /s/ energy in the continuum step as a scaled/centered continuous variable. The results of these models revealed no significant effect of bias in the 1x dose condition ($\beta = 0.586$, $SE = 0.406$, $z = 1.443$, $p = 0.149$) and significant effects of bias in the 4x ($\beta = 1.009$, $SE = 0.407$, $z = 2.477$, $p = 0.013$), 10x ($\beta = 2.498$, $SE = 0.396$, $z = 6.310$, $p < 0.001$), and 20x ($\beta = 3.663$, $SE = 0.448$, $z = 8.182$, $p < 0.001$) dose conditions. Thus, these results indicate that four exposures

[3] In addition to the R packages cited in the main text, we also acknowledge the dplyr and ggplot2 packages from the tidyverse suite (Wickham et al., 2019) that were used for data manipulation and data visualization, and the interactions (Long, 2019) and cowplot (Wilke, 2019) packages that were used for data visualization.

were sufficient to induce a learning effect as indexed by a statistically reliable influence of lexical bias on *asi* responses at test.

To address our primary question, whether the magnitude of the learning effect scales with exposure dose, trial-level *asi* responses from all four dose conditions were examined in a single GLMM. The model included fixed effects of bias, dose, and their interaction. Dose was entered into the model as a series of sliding contrasts that compared consecutive exposure doses (i.e., 1x vs. 4x, 4x vs. 10x, 10x vs. 20x). The random effects structure was identical to that described for the individual dose models. The results revealed a robust effect of bias (β = 2.053, *SE* = 0.212, *z* = 9.681, $p < 0.001$) reflecting more *asi* responses for listeners biased to interpret the ambiguity as /s/ compared to those biased to interpret the ambiguity as /s/. There was no main effect of dose for either the 1x vs. 4x (β = -0.046, *SE* = 0.285, *z* = -0.161, $p < 0.872$) or the 4x vs. 10x (β = 0.425, *SE* = 0.288, *z* = 1.477, *p* = 0.140) contrasts; however, there were more *asi* responses in the 20x compared to the 10x condition (β = 0.836, *SE* = 0.297, *z* = 2.812, *p* = 0.005). Critically, there was no interaction between bias and dose for the 1x vs. 4x contrast (β = 0.489, *SE* = 0.571, *z* = 0.856, *p* = 0.392), providing no evidence to suggest that the magnitude of the bias effect differed between these two doses. In contrast, a significant bias by dose interaction was observed for the 4x vs. 10x contrast (β = 1.510, *SE* = 0.574, *z* = 2.630, *p* = 0.009) and for the 10x vs. 20x contrast (β = 1.261, *SE* = 0.595, *z* = 2.121, *p* = 0.034). In both cases, the direction of the beta estimate indicates that the bias by dose interaction reflects a larger effect of bias for the higher dose; that is, the bias effect was larger for the 10x compared to the 4x dose condition and for the 20x compared to the 10x dose condition. These interactions support the hypothesis that the magnitude of the learning effect scales with exposure dose.

4. EXPERIMENT 2

Consistent with the predictions of the ideal adapter model, the results of experiment 1 demonstrated that perceptual learning was graded in response to the quantity of the evidence in the input to support belief-updating. Specifically, the magnitude of the learning effect monotonically increased across the 4x, 10x, and 20x dose conditions, suggesting that increased evidence is associated with increased learning. Recall that all dose conditions in experiment 1 presented consistent evidence in support of belief-updating. That is, exposure in all dose conditions provided evidence that the talker consistently produced either the /s/ or /ʃ/ category with "ambiguous" variants. As reviewed in the introduction, Tzeng et al. (2021) observed that learning was attenuated when exposure consisted of 10 ambiguous variants *and* 10 clear productions of the same category compared to when exposure consisted of 20 ambiguous variants. Accordingly, diminished learning could reflect a decreased quantity of evidence or decreased consistency of evidence.

The goal of experiment 2 is to examine how quantity *and* consistency influence lexically guided perceptual learning. To do so, two additional groups of listeners completed a 10x dose condition in which they heard 10 ambiguous variants in either an /s/- or /ʃ/-biasing context along with 10 clear variants of the same category. Compared to the 10x dose condition in experiment 1, the quantity of evidence suggesting that the talker produces the biased category with an atypical phonetic pattern is equated, yet the global context yields inconsistent evidence given that listeners also heard 10 clear productions of the biased category. Performance for the inconsistent 10x dose condition in experiment 2 is compared to the 10x and 20x dose conditions in experiment 1. If quantity of evidence is the key determinant of lexically guided perceptual learning, then the magnitude of learning will be equivalent between the inconsistent and consistent 10x dose conditions, which will both show weaker learning compared to the consistent

20x dose condition. If consistency of evidence further impacts learning, then learning will be weaker in the inconsistent 10x dose condition compared to the consistent 10x dose condition. As shown in Figures 2 and 4, the model simulations demonstrated that the ideal adapter framework predicts that quantity of evidence is the putative factor.

4.1 Methods

4.1.1. *Participants*

A different sample of 100 participants were recruited from the Prolific participant pool following all criteria outlined for experiment 1. The final sample included 41 men and 59 women with a mean age of 25 years (*SD* = 5 years). Participants were randomly assigned to either the /s/-bias or /ʃ/-bias exposure condition, yielding 50 participants in each cell. As described above (and shown in Table 1), dose was held constant across all listeners reflecting 10 ambiguous productions *and* 10 clear productions of the biased category. An additional three participants were tested but excluded from analyses due to failure to pass the headphone screen.

4.1.2. Stimuli

The stimuli were identical to those described for experiment 1.

4.1.3. Procedure

The procedure was identical to that outlined for experiment 1 with one key exception: all listeners heard 20 /s/ words and 20 /ʃ/ words during exposure (along with 60 filler words and 100 nonwords), as shown in Table 1. For listeners in the /s/-bias condition, these words included 10 ambiguous /s/ words, 10 clear /s/ words, and 20 clear /ʃ/ words. For listeners in the /ʃ/-bias condition, these words included 10 ambiguous /ʃ/ words, 10 clear /ʃ/ words, and 20 clear /s/ words.

4.2 Results

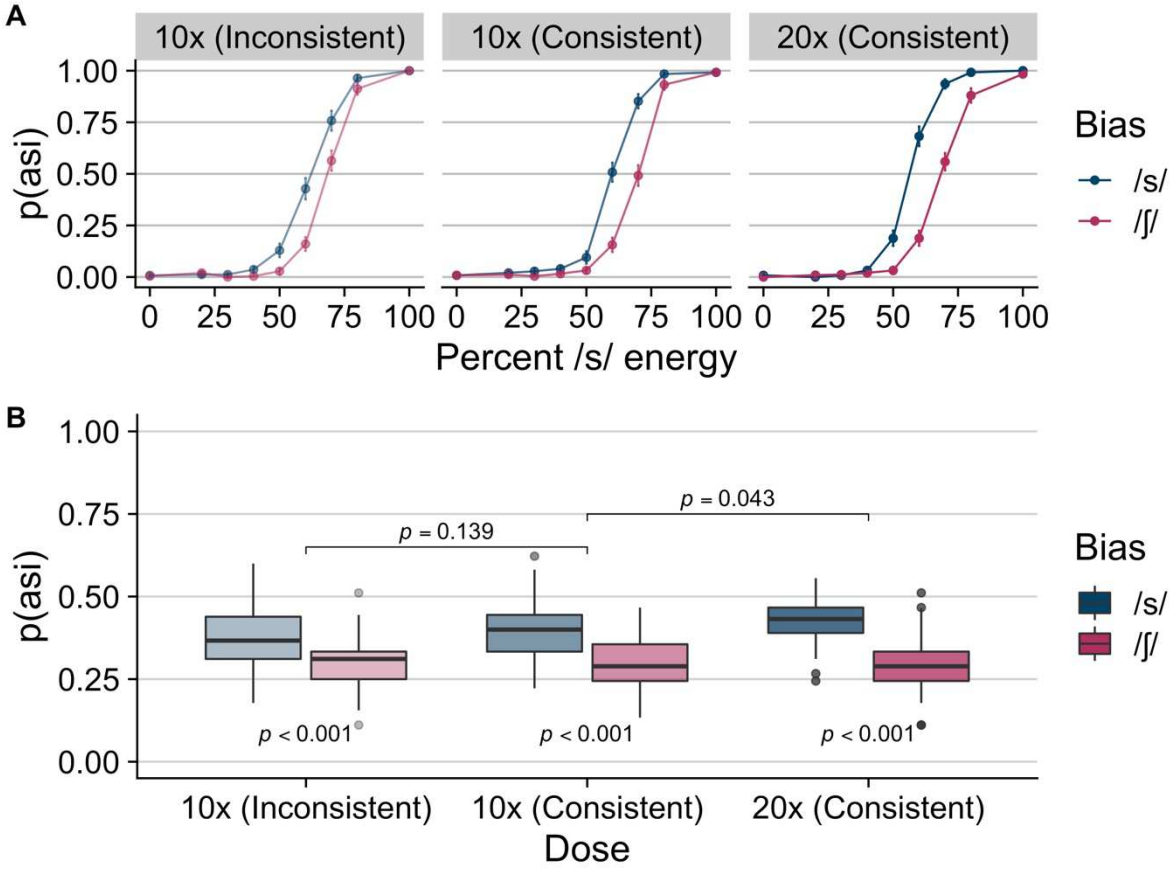
Performance during the exposure phase was analyzed as described for experiment 1. As shown in Table 2, mean lexical decision was near ceiling for each bias condition, as expected given that high accuracy during exposure was an inclusion criterion for participation.[4]

Performance at test was also analyzed as described for experiment 1. Figure 6, panel A shows mean proportion *asi* responses for each continuum step separately for each bias by dose condition; means reflect grand means calculated over by-subject means. This figure shows the inconsistent 10x dose condition (reflecting the new sample tested for experiment 2) along with the consistent 10x and 20x dose conditions (reflecting the samples tested in experiment 1). Figure 6, panel B shows the distribution of by-subject means for each bias by dose condition collapsing across continuum step. Inspection of Figure 6 suggests that the magnitude of the learning effect (i.e., the difference between the two bias conditions) is comparable between the inconsistent 10x and consistent 10x dose conditions, which both show weakened learning compared to the consistent 20x dose condition. To analyze these patterns statistically, we first

[4] As for experiment 1, we conducted a secondary analysis of lexical decision accuracy during the exposure phase that is reported in full in the Supplementary Material. We first examined whether accuracy (percent correct) for each item type differed between the two bias conditions. Accuracy for /s/ words was slightly lower for the /s/-bias condition (*mean* = 96.4, *SD* = 7.0) compared to the /ʃ/-bias condition (*mean* = 99.4, *SD* = 1.9); however, no significant difference between the two bias conditions was observed for the /ʃ/ words. Because listeners in each bias condition of experiment 2 heard *both* ambiguous and natural variants of the to-be-learned category (i.e., /s/ words for the /s/-bias condition, /ʃ/ words for the /ʃ/-bias condition), in contrast to listeners in experiment 1 who only heard ambiguous variants of the to-be-learned category, we also examined lexical decision accuracy between the two variants for each bias condition. Listeners in the /s/-bias condition had slightly lower lexical decision accuracy for the ambiguous variants (*mean* = 93.2, *SD* = 13.9) compared to the clear variants of /s/ words (*mean* = 99.6, *SD* = 2.0). For listeners in the /ʃ/-bias condition, there was no statistically significant difference in lexical decision accuracy between the ambiguous variants (*mean* = 98.3, *SD* = 4.7) and clear variants (*mean* = 98.7, *SD* = 4.1) of /ʃ/ words. As was observed in experiment 1, accuracy in all cases was near ceiling, indicating that the ambiguous variants were endorsed as lexical items. The reader is referred to the Supplementary Material for a comprehensive reporting of this secondary analysis.

tested for a main effect of bias in the inconsistent 10x dose condition. The GLMM revealed a significant effect of bias ($\beta = 1.725$, $SE = 0.451$, $z = 3.882$, $p < 0.001$), reflecting more *asi* responses in the /s/-bias compared to the /ʃ/-bias exposure condition. This result indicates that even in the face of inconsistent evidence regarding a talker's characteristic production, 10 exposures to an "ambiguous" production in a lexically-biased context were sufficient to induce perceptual learning.

Figure 6. Results of experiment 2. Panel A shows mean proportion *asi* responses at test for each continuum step separately for each bias by dose condition. Means reflect grand means calculated over by-subject averages; error bars indicate standard error of the mean. Panel B shows the boxplot distribution of mean proportion *asi* responses across listeners in each bias by dose condition collapsing across continuum step. Annotations below the boxplot distributions indicate *p*-values associated with the main effect of bias in each dose condition. Annotations above the boxplot distributions indicate *p*-values associated with the bias by dose interaction. As described in the main text, the 10x (Inconsistent) condition tested in experiment 2 was compared to the 10x (Consistent) and 20x (Consistent) conditions that were tested in experiment 1, which are plotted together here to promote more direct visual comparison.



Second, we examined the magnitude of the learning effect across dose conditions. As for experiment 1, the GLMM included fixed effects of bias, dose, and their interaction, with the random effects structure including random intercepts by subject and random slopes by subject for continuum step. Dose was entered into the model as a series of sliding contrasts (i.e., 10x inconsistent vs. 10x consistent, 10x consistent vs. 20x consistent). The model revealed a significant effect of bias ($\beta = 2.763$, $SE = 0.253$, $z = 10.940$, $p < 0.001$). The main effect of dose was not reliable for the 10x inconsistent vs. 10x consistent contrast ($\beta = 0.147$, $SE = 0.291$, $z = 0.505$, $p = 0.613$), but it was reliable for the 10x consistent vs. 20x consistent contrast ($\beta = 0.817$, $SE = 0.299$, $z = 2.732$, $p = 0.006$), reflecting more *asi* responses in the latter compared to the former. Critically, the model showed no significant interaction between bias and dose for the 10x conditions ($\beta = 0.863$, $SE = 0.583$, $z = 1.480$, $p = 0.139$), providing no evidence to suggest that the magnitude of the bias effect differed for inconsistent compared to consistent input when dose of ambiguous productions was held constant. As expected, there was a significant interaction between bias and dose for the 10x consistent and 20x consistent dose conditions ($\beta = 1.207$, $SE = 0.598$, $z = 2.020$, $p = 0.043$), reflecting a stronger bias effect in the latter condition.

5. DISCUSSION

There is tremendous variability in how a given linguistic message manifests as acoustic patterns in speech input, including variability that is conditioned on who is speaking. Explicating how listeners solve this lack of invariance problem has been a key goal of speech perception research for decades. Here, we contribute to this effort by examining the degree to which human behavior is predicted by a formal instantiation of the ideal adapter framework for speech adaptation. A core prediction of the ideal adapter framework is that learning should be *graded* to reflect the extent of evidence for adaptation in speech input, yet empirical work has left this

hypothesis largely unacknowledged and not conclusively tested. Conversely, behavioral findings have implicated consistency of evidence as a potential determinant of learning, which extant theoretical models have yet to explore. In the current work, computational modeling of the ideal adapter framework confirmed the key tenet under examination here; specifically, the model simulations predicted that lexically guided perceptual learning would be graded to reflect the quantity of evidence in the input. In contrast, the model simulations predicted no influence of input consistency on lexically guided perceptual learning. The model predictions were largely upheld in human listeners. The results of experiment 1 showed that significant learning was observed given exposure to four or more ambiguous tokens, with the magnitude of learning monotonically increasing given exposure to four, 10, or 20 critical tokens. The results of experiment 2 found no evidence of a differential learning effect given consistent or inconsistent evidence; that is, no reliable difference in learning was observed between listeners who only heard 10 ambiguous productions compared to listeners who heard 10 ambiguous productions and 10 clear productions of the same category.

These findings provide support for the ideal adapter model as a plausible framework for understanding lexically guided perceptual learning specifically and adaptation in speech perception more generally, consistent with recent findings that have invoked the ideal adapter framework as an explanatory theory of perceptual learning (Cummings & Theodore, 2022; Liu & Jaeger, 2018, 2019; Luthra et al., 2021; Saltzman & Myers, 2021; Theodore et al., 2019; Theodore & Monto, 2019; Tzeng et al., 2021). Specifically, modeling behavior as a cumulative integration of local evidence (i.e., input during the exposure phase) with global experience (i.e., prior knowledge) yielded predictions that qualitatively aligned with human performance. Formalizing the theoretical claims of the ideal adapter framework through a computational

implementation of this theory allowed for predictions to be specified beyond a "verbal theory" level of engagement with this framework (van Rooij & Blokpoel, 2020), and theoretical claims may differ based on the level of theory specification. For example, in Tzeng et al. (2019), we hypothesized that the reason why learning was attenuated when listeners heard 10 ambiguous productions along with 10 clear productions of the same category compared to when listeners only heard 20 ambiguous productions of a given category was because learning reflected sensitivity to both the quantity and consistency of evidence for the ambiguous production. However, formal engagement with the ideal adapter theory in the current work revealed that quantity alone was predicted to be the putative factor for adaptation.

Formalizing the theoretical claims of the ideal adapter framework though a computational instantiation also requires critical assumptions to be explicitly identified. The predictions and subsequent interpretation of the current results are of course bound to these assumptions. Here we consider four key assumptions that were made in the current work. First, the ideal adapter framework posits that speech sounds are represented by cue-sound mappings; here we posited that spectral center is a putative cue for the /ʃ/-/s/ contrast. Though spectral center is often used to capture the distinction between /ʃ/ and /s/ (Newman et al., 2001), it is certainly not the only measure disambiguating these two sounds; indeed, there are over 40 acoustic properties that distinguish fricatives from one another (Jongman et al., 2000; McMurray & Jongman, 2011). Many of these cues show high correlations among each other, and so this simplification is perhaps warranted or at least acceptable. Second, our simulations assumed that the "ambiguous" tokens presented during exposure were perfectly intermediate between listeners' prior categories for /ʃ/ and /s/ and, related, that prior experience was constant across listeners. As we discuss further below, this follows convention to use 50/50 blends of /ʃ/ and /s/ energy to create

perceptually ambiguous tokens; however, this convention may not in fact yield acoustically identically variants (given that the precise acoustic pattern is contingent on spectral center of the two clear productions that were used to create the custom 50/50 blend for each word). Moreover, individuals may differ in their stored representations of cue-sound mappings. Third, we assumed that perception of the ambiguous variants was perfectly in line with the intended lexically-biased category. That is, the model simulations assumed perfect identification of the ambiguous variant as /s/ for the /s/-bias simulations and as /ʃ/ for the /ʃ/-bias simulations. This is reasonable given that near ceiling accuracy is observed for the lexical decision exposure task both in the extant literature and in the current work; however, a more accurate assumption may be to introduce a minor level of response noise to simulate accuracy that is not perfectly at ceiling. Fourth, for the simulations presented in the main text, we assumed a moderate level of confidence in prior knowledge. The Supplementary Material present the results of simulations that used lower confidence levels and thus assumed a system with greater flexibility. Though the qualitative patterns regarding the influence of quantity and consistency on the degree of adaptation held across all confidence level parameters, the magnitude of the predicted lexically guided perceptual learning effect was influenced by the prior specification such that larger effects were predicted for lower compared to higher confidence levels. These assumptions appear sufficient for the ideal adapter model to generate predictions that qualitatively align with observed human behavior; however, future research is needed to evaluate each assumption, its predicted influence on learning, and the alignment between theoretical predictions and observed human behavior in greater detail.

As described in the introduction, inspection of learning outcomes in the lexically guided perceptual learning literature reveals wide heterogeneity in the magnitude of adaptation. To date,

the specific magnitude of the learning effect has received little consideration or formal investigation relative to the convention of considering learning as a binary outcome (i.e., present vs. absent, cf. Cummings & Theodore, 2022; Liu & Jaeger, 2018; Tzeng et al., 2021). The ideal adapter model is a theoretical framework that is particularly well-suited for understanding graded learning outcomes that may reflect specific methodological characteristics of a given examination of lexically guided perceptual learning. For example, previous investigations in this domain have examined perceptual learning for different phonetic contrasts, different stimulus sets for a given a phonetic contrast, and different quantities of exposure. The current results confirm that the quantity of exposure can be linked to the magnitude of the learning effect, at least over the range of exposure doses examined here. The ideal adapter framework can also be invoked to examine how stimulus-level factors may contribute to graded learning outcomes. For example, consider a situation in which listeners receive an equal "dose" of atypical input. Holding dose constant, the ideal adapter framework can yield predictions in which learning may differ depending on the specific acoustic-phonetic nature of the ambiguous input. The convention for selecting ambiguous exemplars in this paradigm is via perceptual criteria (i.e., selecting a blend that is deemed perceptually ambiguous, which most often reflects a 50/50 blend of two natural fricative productions), a practice that may not yield equivalent acoustic patterns across stimulus sets. If, for example, the selected "ambiguous" variants exhibit acoustic patterns that are relatively close to prior expectations, then learning would be predicted to be smaller in magnitude compared to ambiguous variants that are perfectly intermediate to the two categories under examination. The ideal adapter framework predicts that learning will reflect the degree to which the atypical input deviates from prior expectations and thus provides a means to link graded learning outcomes to the specific acoustic input presented during the exposure phase.

Indeed, there is some evidence to support stimulus-level influences on the magnitude of lexically guided perceptual learning (Babel et al., 2019). Because formal instantiation of the ideal adapter framework allows simulation of both stimulus-level (i.e., acoustic characteristics of input) and environment-level (i.e., quantity of evidence) factors, this framework could also be used to examine potential interactions between these factors. Recall that Liu and Jaeger (2018) found no significant interaction between the magnitude of the learning effect and exposure dose, in contrast to the robust influence of exposure dose on learning that was observed in the current work. This discrepancy may reflect a lack of power to detect this interaction in Liu and Jaeger (2018), given that their study was not designed to test this question specifically. However, it may also reflect a potential interaction between exposure dose and stimulus-specific properties that may have differed between the two investigations.

Discussion to this point regarding evidence in the input and subsequent incrementality of learning has been limited to the exposure phase in the lexically guided perceptual learning paradigm. This is consistent with the convention to consider the test phase as a measure of learning outcomes that arise given adaptation to the exposure input. As described in the introduction, the ideal adapter framework provides a unifying account of adaptation that occurs in response to explicit supervisory signals, such as disambiguating lexical context, and adaptation that occurs given implicit learning signals, such as unsupervised changes in the distributional patterns in speech input (Kleinschmidt, 2019; Kleinschmidt & Jaeger, 2015, 2016; Theodore et al., 2019; Theodore & Monto, 2019; Tzeng et al., 2021). Recall that the test phase in this paradigm generally consists of a repeated presentation of a uniform continuum of tokens spanning the acoustic space between the sounds of interest, without lexical guidance to label the input as one category versus the other. Recent findings demonstrate that the magnitude of the

lexically guided perceptual learning effect attenuates over longer test phases (Giovannone & Theodore, 2021; Liu & Jaeger, 2019; Scharenborg & Janse, 2013; Tzeng et al., 2021), suggesting that listeners engage in some degree of adaptation to the test stimuli themselves. Attenuation of learning across the test phase is predicted by the ideal adapter framework. Specifically, the posterior beliefs given exposure input continue to be updated given exposure to the test stimuli, which results in an attenuation of the lexically-driven learning effect given experience with the uniform distribution of acoustic-phonetic variants at test. Exploratory analyses in the current study, which compared learning in the primary test phase to learning in a second test phase that was completed directly after the primary test phase, confirm the attenuation of lexically-guided learning over longer test phases. These analyses, along with parallel computational stimulations, are presented in the Supplementary Material.

The atrophy of learning over the course of longer test phases has a relevant implication even beyond providing empirical support for the ideal adapter framework and associated graded learning outcomes. Specifically, the nature of the test phase across the lexically guided perceptual learning literature shows even more heterogeneity than researcher decisions regarding the exposure phase. For example, some test phases have only presented potentially ambiguous variants (e.g., Liu & Jaeger, 2018) while others also included multiple clear continuum endpoints (e.g., Tzeng et al., 2021). The resolution of the test continuum also varies, with some studies presenting six steps that span clear endpoints (e.g., Kraljic et al., 2008) and other studies presenting as many as 60 unique test tokens (e.g., Chládková et al., 2017). Furthermore, the duration of the test phase varies, with some studies presenting a limited number of test cycles resulting in 30 test trials (e.g., Eisner & McQueen, 2005) and other studies presenting extensive test cycles of multiple test continua leading to 480 total test trials (e.g., Kraljic & Samuel, 2006).

The ideal adapter models predicts that the conditions of the test phase – including the specific acoustic characteristics of the test stimuli and the quantity of input received during test (i.e., the duration of the test phase) – will itself impact performance during test and thus influence the degree to which the test phase provides a "pure" measure of adaptation that arises given lexically-biased input during the exposure phase. Accordingly, the framework provides a means to test the hypothesis that heterogeneity of learning outcomes in the extant literature may reflect graded learning outcomes that are linked not only to the specific conditions of the critical exposure phase, but also to the specific conditions of the test phase.

Operationalizing adaptation as a graded outcome that reflects a continuous integration of observed evidence with beliefs formed over long-term experience – consistent with the primary tenet of the ideal adapter framework – holds extreme promise for advancing theoretical accounts of perceptual learning. This viewpoint also has the potential to reconcile what on the surface may appear to be discrepant findings in the literature. As described in the introduction, two recent investigations examined the influence of exposure dose on perceptual learning for multiple talkers. Luthra et al. (2021) observed no statistically significant learning effect given exposure to 16 critical productions, which was observed given exposure to 32 critical productions. Following the convention to interpret learning as a binary outcome, the conclusion was that adaptation to multiple talkers required twice the conventional exposure dose, despite the failure to observe a significant interaction between dose and learning. Cummings and Theodore (2022) found evidence of perceptual learning for multiple talkers with exposure doses of both 20 and 40 critical productions, and – as in Luthra et al. (2021) – found no evidence of an interaction between dose and learning, suggesting a ceiling effect on learning consistent with 20 exposures providing sufficient evidence to fully accommodate the atypical production. At first blush, these

discrepant results present a challenge: in terms of a learning binary, does perceptual learning for multiple talkers require additional exposure? This apparent conflict resolves itself when viewed through the lens of graded learning. Both studies found no evidence of an interaction between dose and learning. However, in Luthra et al. (2021), learning effects within each dose condition happen to straddle the threshold for statistical significance, while in Cummings and Theodore (2022), learning effects in both conditions landed on the same side of this threshold. Furthermore, formal comparisons of the effect sizes observed in these two studies revealed that the magnitude of the learning effect across all conditions of Luthra et al. (2021) were smaller than those of Cummings and Theodore (2022); most notably, the magnitude of the learning effect for the standard dose condition in Cummings and Theodore was approximately twice as large as the magnitude of the double dose condition in Luthra et al. (2021). These two studies illustrate the importance and promise for considering learning as a graded outcome that may reflect the specific to-be-learned acoustic characteristics of a given stimulus set, which is not optimally captured when learning is considered as a binary outcome.

Indeed, we submit that after a foundational 20 years of incredibly fruitful investigation of lexically guided perceptual learning, future work stands to benefit from a paradigm shift to consider learning outcomes beyond the binary. The ideal adapter model, which has increasingly been invoked as an explanatory theory of perceptual learning for speech, provides a theoretical framework for examining learning as a graded outcome in response to numerous aspects of the listening environment. The current work provides some evidence in support of its primary tenet; learning reflects the *incremental* updating of cue-sound mappings to incorporate observed evidence with prior beliefs. Leveraging the computational instantiation of this framework would support refinement of researcher assumptions that are critical for understanding how behavioral

patterns are linked to this theory. Additionally, it potentially provides a means for generating fine-grained predictions of human behavior that reflect specific aspects of experimental tasks including idiosyncratic aspects of to-be-learned stimuli. Despite the promising support for the ideal adapter model as a theory of speech adaptation (Cummings & Theodore, 2022; Kleinschmidt, 2019, 2019; Kleinschmidt & Jaeger, 2015, 2016; Liu & Jaeger, 2018; Luthra et al., 2021; Theodore et al., 2019; Theodore & Monto, 2019; Tzeng et al., 2021), future research is needed to determine whether this theory is in fact sufficient to explain the wide variability in learning outcomes that is observed across the lexically guided perceptual learning literature.

6. CONCLUSION

The current investigation marries a computational instantiation of the ideal adapter framework with a behavioral investigation of lexically guided perceptual learning. This union confirms a primary tenet of the ideal adapter framework and, in turn, establishes quantity of evidence as a key determinant of adaptation in human listeners. Specifically, we found evidence that lexically guided perceptual learning is not a binary outcome; rather, learning is graded in response to the quantity of evidence in the input. Moreover, we found no evidence to suggest that learning was linked to consistency of the evidence, in line with the predictions of this computational framework. The current results provide a strong foundation for future research that links graded learning outcomes to other aspects of the input, including the specific acoustic instantiation of to-be-learned input. We submit that future work, both behavioral and theoretical, will benefit from a departure from the convention to consider perceptual learning as a binary outcome. Continued investigation of graded learning outcomes in response to variability in speech input will be maximally fruitful for explicating the mechanisms that allow listeners to dynamically modify speech perception to reflect structured regularities in speech input.

Acknowledgments

This research was supported by National Science Foundation (BCS) grant 1827591 to RMT. SNC was supported by the Donald Shankweiler Language Sciences Award (University of Connecticut). The views expressed here reflect those of the authors and not the National Science Foundation.

References

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

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388-4071–20. https://doi.org/10.3758/s13428-019-01237-x

Babel, M., McAuliffe, M., Norton, C., Senior, B., & Vaughn, C. (2019). The Goldilocks zone of perceptual learning. *Phonetica*, *76*(2–3), 179–200.

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 auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*(6), 592–597.

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

Byrd, D. (1992). Preliminary results on speaker-dependent variation in the TIMIT database. *The Journal of the Acoustical Society of America*, *92*(1), 593–596.

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.

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, *61*, 30–47.

Chodroff, E., & Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard*, *4*(s2), 20170047.

Cummings, S., N., & Theodore, R. M. (2022). Perceptual learning of multiple talkers: Determinants, characteristics, and limitations. *Attention, Perception & Psychophysics*, *84*, 2335–2359.

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

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

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238.

Fant, G. (1973). *Speech sounds and features*. MIT Press.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110–125.

Giovannone, N., & Theodore, R. M. (2021). Individual differences in lexical contributions to speech perception. *Journal of Speech, Language, and Hearing Research*, *64*(3), 707–724.

Heffner, C. C., Fuhrmeister, P., Luthra, S., Mechtenberg, H., Saltzman, D., & Myers, E. B. (2022). Reliability and validity for perceptual flexibility in speech. *Brain and Language*, *226*, 105070.

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

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1009–1021.

Idemaru, K., & Holt, L. L. (2020). Generalization of dimension-based statistical learning. *Attention, Perception, & Psychophysics*, *82*(4), 1744–1762.

Jesse, A. (2021). Sentence context guides phonetic retuning to speaker idiosyncrasies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(1), 184–194. https://doi.org/10.1037/xlm0000805

Johnson, K., & Beckman, M. E. (1997). Production and perception of individual speaking styles. In *Working Papers in Linguistics* (Vol. 50, pp. 115–125). Ohio State University, Department of Linguistics.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252–1263.

Keetels, M., Schakel, L., Bonte, M., & Vroomen, J. (2016). Phonetic recalibration of speech by text. *Attention, Perception, & Psychophysics*, *78*(3), 938–945.

Klatt, D. H. (1986). The problem of variability in speech recognition and in models of speech perception. In *Invariance and variability in speech processes* (pp. 301–324). Erlbaum.

Kleinschmidt, D. F. (2017). *beliefupdatr: Belief updating for phonetic adaptation.* (R package version 0.0.3). https://github.com/kleinschmidt/beliefupdatr

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.

Kleinschmidt, D. F., & Jaeger, T. F. (2016). What do you expect from an unfamiliar talker? *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*.

Kluender, K. R., Stilp, C. E., & Lucas, F. L. (2019). Long-standing problems in speech perception dissolve within an information-theoretic perspective. *Attention, Perception, & Psychophysics*, *81*(4), 861–883.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268.

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.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

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

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

Liu, L., & Jaeger, T. F. (2019). Talker-specific pronunciation or speech error? Discounting (or not) atypical pronunciations during speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(12), 1562–1588.

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1783–1798.

Long, J. A. (2019). *interactions: Comprehensive, User-Friendly Toolkit for Probing Interactions* (R package version 1.0.0). https://cran.r-project.org/package=interactions

Luthra, S., Mechtenberg, H., & Myers, E. B. (2021). Perceptual learning of multiple talkers requires additional exposure. *Attention, Perception, & Psychophysics*, *83*, 2217–2228.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, *12*(3), 369–378.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219–246.

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

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, *53*(4), 1551–1562.

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*(2), 356–361.

Monto, N. R. (2018). *SlopeExtractR: Creates identification functions and extracts slopes from belief-updated normal distributions.* (R package version 0.0.0.9). https://github.com/nick-monto/slopeExtractR

Munson, B. (2011). The influence of actual and imputed talker gender on fricative perception, revisited. *The Journal of the Acoustical Society of America*, *130*(5), 2631–2634.

Nelson, S., & Durvasula, K. (2021). Lexically-guided perceptual learning does generalize to new phonetic contexts. *Journal of Phonetics*, *84*, 101019.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181–1196.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Nygaard, L. C., & Lunders, E. R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory & Cognition*, *30*(4), 583–593.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Attention, Perception, & Psychophysics*, *57*(7), 989–1001.

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

Reinisch, E., & Mitterer, H. (2016). Exposure modality, input variability and the categories of perceptual recalibration. *Journal of Phonetics*, *55*, 96–108.

Saltzman, D., & Myers, E. (2021). Listeners are initially flexible in updating phonetic beliefs over time. *Psychonomic Bulletin & Review*, 1–11.

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. https://doi.org/10.1016/j.cogpsych.2016.06.007

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*(6), 1207–1218. https://doi.org/10.3758/APP.71.6.1207

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

Schuhmann, K. S. (2012). Perceptual learning in Hindi-English bilinguals. *Studies in the Linguistic Sciences: Illinois Working Papers*, 81–99.

Sidaras, S. K., Alexander, J. E., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America*, *125*(5), 3306–3316.

Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *The Journal of the Acoustical Society of America*, *128*(4), 2090–2099.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America*, *125*(6), 3974–3982. https://doi.org/10.1121/1.3106131

Theodore, R. M., & Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions. *Psychonomic Bulletin & Review*, *26*(3), 985–992.

Theodore, R. M., Monto, N. R., & Graham, S. (2019). Individual differences in distributional learning for speech: What's ideal for ideal observers? *Journal of Speech, Language, and Hearing Research*, 1–13.

Tzeng, C. Y., Nygaard, L. C., & Theodore, R. M. (2021). A second chance for a first impression: Sensitivity to cumulative input statistics for lexically guided perceptual learning. *Psychonomic Bulletin & Review*, *28*, 1003–1014.

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. https://doi.org/10.1037/0096-1523.33.6.1483

van Rooij, I., & Blokpoel, M. (2020). Formalizing verbal theories: A tutorial by dialogue. *Social Psychology*, *51*(5), 285–298.

Weatherholtz, K., & Jaeger, T. F. (2016). Speech perception and generalization across talkers and accents. In *Oxford Research Encyclopedia of Linguistics*.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

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

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

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America*, *143*(4), 2013–2031.

Zheng, Y., & Samuel, A. G. (2020). The relationship between phonemic category boundary changes and perceptual adjustments to natural accents. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(7), 1270–1292.