# 1 Abstract

Human speech perception is a computational feat. Now recognized as critical to this human ability is *adaptivity*: a few minutes of exposure can significantly reduce the processing difficulty listeners experience during initial encounters with an unfamiliar accent. How such adaptation unfolds incrementally, however, remains largely unknown, leaving basic predictions by theories of adaptive speech perception untested. This includes questions about how listeners' prior expectations based on lifelong experiences are integrated with the unfamiliar speech input, as well as questions about the speed and success of adaptation. We begin to address these knowledge gaps in a novel incremental exposure-test paradigm. We expose US English listeners to shifted phonetic distributions of word-initial stops (e.g., "dill" vs. "till"), while incrementally assessing cumulative changes in listeners' perception. We use Bayesian mixed-effects psychometric models to characterize these changes, and compare listeners' behavior against both idealized learners (ideal observers that know the exposure statistics) and a model of adaptive speech perception (ideal adaptors that have to infer those statistics). Our findings support several previously untested predictions of distributional learning models of adaptive speech perception. Our findings do, however, also suggest previously unrecognized limits on adapti10/20vity that are unexpected under *any* existing model.

*Keywords:* speech perception; adaptation; incremental; distributional learning; error-driven learning; ideal adaptor

Word count: X

# 2 Learning to understand an unfamiliar talker: Testing models of adaptive speech perception

## 1 Introduction

Human speech perception is a remarkable feat. Successful speech recognition requires that listeners map the acoustic signal onto words and meanings. But this signal-to-meaning mapping varies across talkers and context. The same word spoken by different talkers can sound quite different; and conversely, the same acoustic signal can imply different words depending on the talker. Yet, listeners typically recognize speech quickly and accurately across a wide range of talkers and acoustic conditions (after decades of advances, automatic speech recognition is just beginning to approach the recognition accuracy that most listeners display during everyday speech perception).

Research has identified *adaptivity* as a key component to the robustness of human speech perception. Although first encounters with an unfamiliar accent can cause initial processing difficulty, this difficulty diminishes with exposure, sometimes rapidly (Bradlow, Bassard, & Paller, 2023; e.g., Bradlow & Bent, 2008; Sidaras, Alexander, & Nygaard, 2009; Xie, Liu, & Jaeger, 2021). Eighteen short sentences from a talker with an unfamiliar accent—even a moderately strong second language accent—have been found to significantly improve subsequent perception of that talker's speech (Clarke & Garrett, 2004; Xie, Weatherholtz, et al., 2018). Findings like these suggest that speech perception can be highly malleable, allowing listeners to adjust the mapping from acoustics to phonetic categories and word meanings. While such adaptivity may seem obvious in hindsight, its discovery was a major breakthrough in the field of speech perception, spurring the development of new paradigms and theories (for reviews, see Bent & Baese-Berk, 2021; Cummings & Theodore, 2023; Schertz & Clare, 2020; Zheng & Samuel, 2023).

We thus know *that* listeners adapt to unfamiliar talkers. What remains unclear, however, is *how* adaptation is achieved. How do listeners integrate information from a new talker, and how does this come to incrementally change their interpretation of that talker's speech? Research on adaptive speech perception tends to discuss informal—often descriptive, rather than

explanatory—hypotheses (see also Norris & Cutler, 2021). This includes references to "boundary re-tuning/shift", "perceptual/phonetic recalibration/retuning", "category shift/expansion" or similar ideas (e.g., McQueen, Cutler, & Norris, 2006; Mitterer, Scharenborg, & McQueen, 2013; Reinisch & Holt, 2014; Schmale, Cristia, & Seidl, 2012; Vroomen & Baart, 2009; Xie, Theodore, & Myers, 2017; Zheng & Samuel, 2020). Such hypotheses do not specify what mechanisms support adaptive speech perception, nor do they make predictions about how adaptation unfolds incrementally with each new observation from an unfamiliar talker.

Viewed from this perspective, research on adaptive speech perception can appear to be an open-ended list of empirical questions. Is adaptation more or less immediate, or does it unfold gradually? If the latter, do changes in listeners' behavior accumulate additively, leading to more or less linear changes in behavior? If not a linear development, is adaptation first slow and then fast, or first fast and then slow? And, how do differences in listeners' prior experience affect how listeners adapt? Are there limits to listeners' ability to fully adapt to a new talker? Or can we adapt to more or less any accent provided sufficient input?

Critically, some existing theories do make clear, quantifiable predictions about all of these questions—including many basic predictions that remain largely untested. One of the most developed of these theories are *distributional learning* models (e.g., Apfelbaum & McMurray, 2015; Harmon, Idemaru, & Kapatsinski, 2019; Johnson, 1997; Kleinschmidt & Jaeger, 2015; Lancia & Winter, 2013; Magnuson et al., 2020; Nearey & Assmann, 2007; Sohoglu & Davis, 2016). These models make untested predictions about how adaptation unfolds incrementally during the initial moments of listening to an unfamiliar talker. While distributional learning models differ from each other in important aspects, they share the central assumption that listeners incrementally learn and store information about talkers' speech. This includes information about the phonetic distributions that characterize the talker's speech, such as the average values of phonetic cues, their variability, or even the full phonetic distributions of all speech categories. These statistical properties are then used to interpret subsequent speech from the talker, supporting robust speech recognition across talkers (for reviews, see Bent & Baese-Berk, 2021; Schertz & Clare, 2020; Xie, Jaeger, & Kurumada, 2023).

With this shared assumption, distributional learning models also share several critical

predictions. Consider a listener's initial encounter with an unfamiliar talker who produces some sounds in an unexpected way (Figure 1A). Listeners' perception is predicted to change incrementally with exposure (Figure 1B-C). Distributional learning models make four predictions about how these changes unfold incrementally. First, the direction and magnitude of that change should gradiently depend on listeners' prior expectations based on relevant previously experienced speech input from other talkers (**prediction 1 - *prior expectations***), and both the amount (**prediction 2a - *exposure amount***) and distribution of phonetic cues in the exposure input from the unfamiliar talker (**prediction 2b - *exposure distribution*** , for review, see Xie et al., 2023). Specifically, listeners' categorization functions—the mapping from acoustics to phonetic categories and words—should gradually shift from a starting point that reflects the statistics of previously experienced speech towards a target that reflects the statistics of the new talker's speech. Existing models further predict that this shift proceeds until the listener has fully learned the statistics of the new talker's speech (**prediction 3 - *learn to convergence***). Finally, some distributional learning models further commit to specific learning mechanisms that constrain how exactly adaptation is expected to accumulate incrementally: both error-driven theories (Harmon et al., 2019; Olejarczuk, Kapatsinski, & Baayen, 2018; Sohoglu & Davis, 2016) and theories of ideal information integration (Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2015) predict that adaptation initially proceeds quickly and then slows down as the listener approaches the correct mapping from the acoustic signal to phonetic categories (**prediction 4 - *diminishing returns***).[1] Figure 1D illustrates these predictions and contrasts them with other possible scenarios.

[1] Predictions (1)-(4) assume that listeners *know* that they are listening to the same new talker. Talker recognition is itself an active inference process that we do not further discuss here (but see Kleinschmidt & Jaeger, 2015; Magnuson & Nusbaum, 2007).

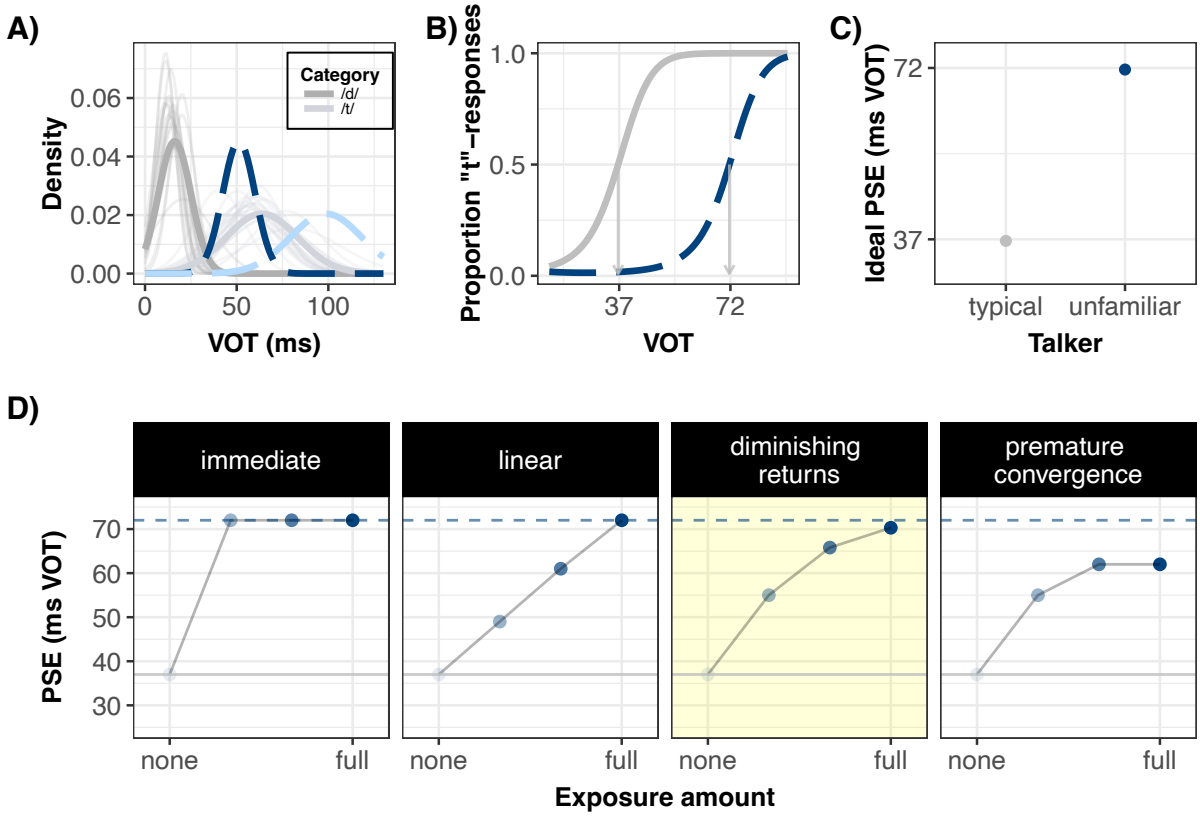


*Figure 1.* Some hypothetical ways in which adaptive changes in listeners perception might unfold incrementally, using the pronunciation of US English word-initial /d/ and /t/ as an example (as in "dip" vs. "tip"). **Panel A:** Transparent lines indicate cross-talker variability in the realization of /d/ and /t/ along the primary cue used to distinguish them (voice onset timing or VOT). Shown are 20 random talkers from a database of connected speech (Chodroff & Wilson, 2018). The thicker solid lines indicate a 'typical' talker (averaging over all talkers in the database). Dashed lines indicate a hypothetical unfamiliar talker with a noticeably different distribution of VOT values. **Panel B:** Ideal categorization functions along the phonetic VOT continuum for speech from a typical talker (*idealized pre-exposure listener*, SI §4.1) and speech from the unfamiliar talker (*idealized learner* that has fully learned that talkers distributions, SI §4.2). Grey arrows point to the points of subjective equality (PSE), the point along the phonetic VOT continuum at which listeners are equally likely to identify a sound as an instance of /d/ or /t/. **Panel C:** The same as in Panel B but just showing the PSE. **Panel D:** Different ways in which listeners' PSEs along the phonetic VOT continuum might incrementally change with increasing exposure to the unfamiliar talker (from more transparent to less transparent). The horizontal lines indicates the ideal PSEs from Panel C.

Although predictions (1)-(4) are specific and testable, they remain largely untested. This is in part due to limitations of the designs and paradigms used in research on adaptive speech perception. The most common paradigms expose one group of listeners to one speech pattern

(e.g., rightward shifted VOT distributions as in Figure 1A), and a second group of listeners to another speech pattern (e.g., leftward shifted VOT distributions). Following exposure, both groups are tested on their ability to recognize one of the two speech patterns. Such designs were effective in establishing the *existence* of adaptive speech perception (see also Cummings & Theodore, 2023). They do, however, offer only weak tests of existing theories (for demonstration, see Xie et al., 2023). Put simply, it is one thing to show that differences in exposure lead to differences in behavior; it is another thing to test whether the direction and magnitude of changes in behavior can be consistently explained by existing theories.

Recent reviews have thus called for the development of paradigms that can more strongly constrain theories of adaptive speech perception (Bent & Baese-Berk, 2021; Schertz & Clare, 2020; Xie et al., 2023). Based on computational simulations, Xie and colleagues argue that strong tests require (a) information about the distribution of phonetic cues in both listeners' prior experience and during exposure, (b) paradigms that measure incremental changes in listeners' behavior both within and across exposure conditions, and (c) analyses that link the latter to the former. The present study responds to this call. We present a novel incremental exposure-test paradigm, and use it to test predictions (1)-(4), repeated in Table 1.

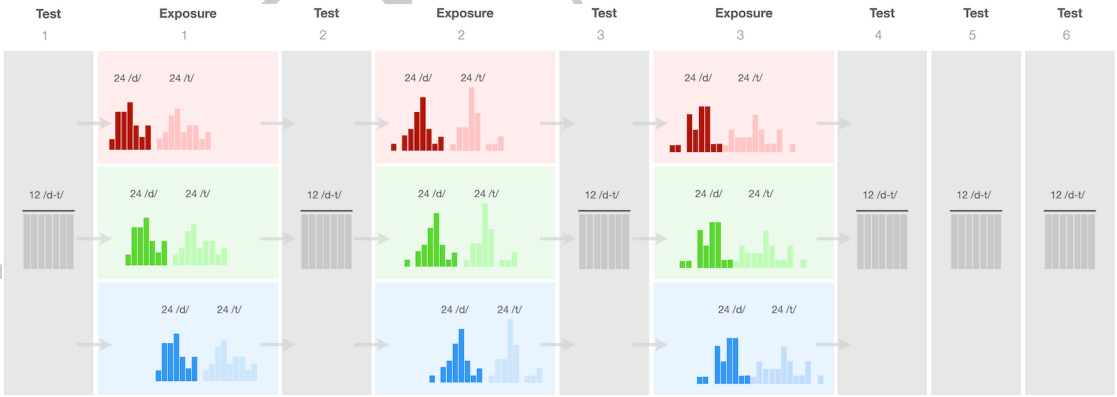


*Figure 2.* Incremental exposure-test design of our experiment. The three exposure conditions (rows) differed in the distribution of voice onset time (VOT), the primary phonetic cue to syllable-initial /d/ and /t/ in English (e.g., "dip" vs. "tip"). Test blocks assessed L1-US English listeners' categorization functions over VOT stimuli that were held identical within and across conditions.

Figure 2 illustrates our approach. Between groups of participants, we manipulate the amount and distribution of phonetic cues in the exposure input. The three exposure distributions

we use are shifted to different degrees both relative to each other, and relative to listeners' prior expectations. This allows us to test predictions (1) and (2a,b) that direction and magnitude of that change should gradiently depend on how and how much the current talker's speech deviates from the listeners' prior expectations. We measure listeners' categorization functions at multiple points during exposure, and determine whether the direction and magnitude of the observed changes in behavior are consistent with the predictions of distributional learning models, including prediction (4) about the diminishing rate of changes in listeners' behavior. To further guide the interpretation of results, we use normative models of adaptive speech perception (ideal observers and adaptors, Feldman, Griffiths, & Morgan, 2009; Kleinschmidt & Jaeger, 2015; Massaro, 1989; Xie et al., 2023). This enables predictions about—intentionally idealized—listeners and distributional learners, prior to considerations about memory or other cognitive limitations. Comparisons of participants' categorization functions against these normative models provides a principled and informative approach to identifying constraints on adaptive speech perception, addressing prediction (3) about learning to convergence.

| Prediction | Evidence that the *outcome* of learning is compatible with this prediction |
|---|---|
| (1) - *prior expectations* | (Kang & Schertz, 2021; Schertz et al., 2016; Tan et al., 2021; Xie et al., 2021) |
| (2a) - *exposure amount* | (Vroomen et al., 2007; Cummings & Theodore, 2023; Kleinschmidt & Jaeger, 2011; Liu & Jaeger, 2018) |
| (2b) - *exposure distribution* | (Chládková et al., 2017; Clayards et al., 2008; Colby et al., 2018; Hitczenko & Feldman, 2016; Idemaru & Holt, 2011; Kleinschmidt, 2020; Theodore & Monto, 2019) |
| (3) - *learn to convergence* | — |
| (4) - *diminishing returns* | — |

Table 1
*Predictions of distributional learning models about incremental adaptation to an unfamiliar talker. To the best of our knowledge, only prediction (2a) has been tested against* incremental *changes in listeners' behavior, and predictions (3) and (4) have not been tested at all.*

Our paradigm integrates, and builds on, advances in separate lines of research on unsupervised distributional learning during speech perception (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Colby, Clayards, & Baum, 2018; Kleinschmidt, 2020; Theodore & Monto, 2019), lexically- or visually-guided perceptual learning (Cummings & Theodore, 2023; Kleinschmidt & Jaeger, 2012; Vroomen, Linden, De Gelder, & Bertelson, 2007), and adaptation to natural accents (Hitczenko & Feldman, 2016; Tan, Xie, & Jaeger, 2021; Xie, Buxó-Lugo, & Kurumada, 2021). In

the general discussion, we return to these and related works in more detail. For now, we make three observations about previous work that motivated our approach.

First, research on adaptation to natural accents does not typically investigate how the phonetic properties of the exposure input relate to changes in listeners' behavior. This is in part due to the methodological challenges inherent to data that are high in ecological validity, but also of high dimensionality: it is simply more complicated to model the acoustic consequences of natural accents. Even notable exception to this trend have thus mostly been limited to broad qualitative comparisons (e.g., Schertz, Cho, Lotto, & Warner, 2016; Xie et al., 2017; see also, Schertz & Clare, 2020), leaving open whether the direction and magnitude of changes in listeners' behavior can be predicted by existing models (but see Hitczenko & Feldman, 2016; Tan et al., 2021; Xie, Buxó-Lugo, et al., 2021). This limitation is generally shared with research on lexically- or visually-guided perceptual learning. Tests of distributional learning models have thus largely relied on paradigms that afford researchers with fine-grained control over the distribution of phonetic properties that listeners experience in the experiment (e.g., Chládková, Podlipskỳ, & Chionidou, 2017; Clayards et al., 2008; Colby et al., 2018; Idemaru & Holt, 2011; Kleinschmidt, 2020; Theodore & Monto, 2019). We follow this approach here. We do, however, take several modest steps towards addressing concerns about the ecological validity of such approaches. This includes concerns about the ecological validity of both the speech stimuli and their distribution in the experiment (see discussion in Baese-Berk, 2018). For example, previous distributional learning studies have often used highly unnatural, 'robotic'-sounding, speech. Beyond raising questions about what types of expectations listeners apply to such speech, these stimuli also failed to exhibit naturally occurring covariation between phonetic cues that listeners are known to expect (see, e.g., Idemaru & Holt, 2011; Schertz et al., 2016). We instead developed stimuli that both sound natural and exhibit the type of phonetic covariation that listeners expect from everyday speech perception. We return to these and additional steps we took to increase the ecological validity of the exposure *distributions* under Methods.

Second, the few studies that have tested predictions of existing models have investigated the *outcome* of learning, leaving open whether adaptive speech perception unfolds over time in ways consistent with distributional learning models. For example, in an important early study,

Clayards et al. (2008) exposed two different groups of US English listeners to instances of "b" and "p" that differed in their distribution along the voice onset time continuum (VOT). VOT is the primary phonetic cue to word-initial stops in US English: the voiced category (e.g., /b/, /d/, or /g/) is produced with lower VOT than the voiceless category (/p/, /t/, /k/). Clayards and colleagues held the VOT means of /b/ and /p/ constant between the two exposure groups, but manipulated whether both /b/ and /p/ had wide or narrow variance along VOT. Using a distributional learning model similar to the idealized learners we presented below, Clayards and colleagues predicted that listeners in the wide variance group would exhibit a more shallow categorization function than the narrow variance group. This is precisely what they found, providing support for prediction (2b) that the distribution of phonetic cues in the exposure input causes changes in listeners' behavior (see also Nixon, Rij, Mok, Baayen, & Chen, 2016; Theodore & Monto, 2019). Findings like these suggests that the outcome of adaptation is qualitatively compatible with predictions (2a) and (2b) of distributional learning models (see also Hitczenko & Feldman, 2016; Tan et al., 2021; Xie, Buxó-Lugo, et al., 2021). Previous studies have, however, relied on tests that averaged over, and/or followed, hundreds of exposure trials.[2] This leaves open whether listeners' categorization behavior follows the change pattern predicted by models of adaptive speech perception: where categorization first reflects expectations based on previously experienced phonetic distributions (prediction 1) and with increasing exposure, integrates the phonetic distributions of the input from the unfamiliar talker (predictions 2a,b). Previous studies also focused on one prediction at a time, leaving open how the effect of prior expectations and the statistics of the unfamiliar input *jointly* explain adaptation.

Third and finally, we are not aware of any previous tests of predictions (3 - *learn to convergence*) and (4 - *diminishing returns*): without incremental testing it is difficult to assess whether there are hard limits on adaptation or simply 'how far the learner has gotten' with the exposure input they have received so far (for discussion, see Cummings & Theodore, 2023; Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016). For the same reasons, it is difficult to assess

[2] A related line of work has used distributional learning or explicit training paradigms to study the acquisition of *novel* sound contrasts (e.g., Maye, Werker, & Gerken, 2002; McClelland, Thomas, McCandliss, & Fiez, 1999; Pajak & Levy, 2012; Pisoni, Aslin, Perey, & Hennessy, 1982). These studies, too, have observed outcomes predicted by distributional learning models (for review, see Pajak, Fine, Kleinschmidt, & Jaeger, 2016), but have left untested the incremental unfolding of learning, or assessed it at time scales much longer than the ones tested here.

whether the build-up of adaptation follows the predictions of error-driven learning or ideal information integration (prediction 4).

The incremental exposure-test paradigm in Figure 2 begins to address these knowledge gaps. To anticipate our results, we find that the changes in listeners' categorization behavior *largely* follow the predictions of distributional learning models. In particular, we present the first direct evidence that the direction and magnitude of changes in listeners' categorization functions is jointly determined by their prior expectations (prediction 1) and the amount and distribution of phonetic cues in the exposure input (predictions 2a,b). We also find initial—though not decisive—evidence that changes in rate of adaptation across exposure are consistent with the predictions of error-driven learning theories and theories of ideal information integration (prediction 4). We show that a Bayesian model of adaptation that is based on principles of ideal information integration (the ideal adaptor, Kleinschmidt & Jaeger, 2015, 2016) predicts participants' responses with very high accuracy ($R^2 = 97\%$). However, not all observations we make are predicted by existing models, providing new insights into previously unrecognized limits of adaptation. In particular, we find little support for prediction (3 - *learn to convergence*). Rather, changes in listeners' behavior seem to plateau long before listeners achieve the categorization functions and accuracy that would be expected if they fully learned the talkers' phonetic distributions (cf. the *premature convergence* panel of Figure 1C). We also find that this constraint on adaptation seems to be asymmetric, depending on the direction of the shift in the exposure input relative to listeners' prior expectations. We discuss the implications of our findings for theories of adaptive speech perception, and suggest how future variants of our paradigm can be used to further contrast different models of adaptive speech perception.

## 1.1 Open science

All data and code for this article are available on OSF at https://osf.io/hxcy4/. Following Xie et al. (2023), both this article and its supplementary information (SI) are written in R Markdown. This allows other researchers to replicate and revise our analyses with the press of a button using freely available software (R, R Core Team, 2022; RStudio Team, 2020, see also SI, §1).

This study was not publicly pre-registered. The design, participant recruitment, and

procedure were internally pre-registered as part of an undergraduate class at the University of Rochester (BCS206/207). The experiment was originally designed to address predictions (1)-(3). Our analyses of prediction (4 - *diminishing returns*) are thus post-hoc, as are some of the analyses we present to understand the evidence against prediction (3 - *learn to convergence*). All post-hoc analyses are indicated as such. Finally, the ideal observer and adaptor models introduced below to guide interpretation of results follow our previous work (Kleinschmidt & Jaeger, 2015; Tan et al., 2021; Xie et al., 2023). However, the choice of phonetic data on which these models are trained constitute researcher degrees of freedom. Where relevant, we motivate our decisions.

# 2 Methods

## 2.1 Participants

We recruited 126 participants from the Prolific crowdsourcing platform. Participants were randomly assigned to one of three exposure conditions in Figure 2. We used Prolific's pre-screening to limit the experiment to participants (1) of US nationality, (2) who reported to be English speaking monolinguals, and (3) had not previously participated in any experiment from our lab on Prolific. Prior to the start of the experiment, participants had to confirm that they (4) had spent the first 10 years of their life in the US, (5) were in a quiet place and free from distractions, and (6) wore in-ear or over-the-ears headphones that cost at least $15.

Participants' responses were collected via Javascript developed by the Human Language Processing Lab at the University of Rochester (Kleinschmidt et al., 2021) and stored via Proliferate developed at, and hosted by, the ALPs lab at Stanford University (Schuster, 2020). Participants took an average of 31.6 minutes (SD = 20 minutes) to complete the experiment and were remunerated $8.00/hour. An optional post-experiment survey recorded participant demographics using NIH prescribed categories, including participant sex (female: 59, male: 60, declined to report: 3), age (mean = 38 years; SD = 12; 95% quantiles = 20-62.1 years), ethnicity (Non-Hispanic: 113, Hispanic: 6, declined to report: 3), and race (due to a technical error, all information was lost).

## 2.2 Materials

We recorded 8 tokens each of four minimal word pairs with word-initial /d/-/t/ (*dill/till*, *dim/tim*, *din/tin*, and *dip/tip*) from a 23-year-old, female L1-US English talker from New Hampshire. In addition to these critical minimal pairs we also recorded three words that did not contain any stop consonant sounds ("flare", "share", and "rare"). These word recordings were used for catch trials. Stimulus intensity was normalized to 70 dB sound pressure level for all recordings.

The critical minimal pair recordings were used to create four VOT continua ranging from -100 to +130 ms in 5 ms steps.[3] Continua were generated using a script (Winn, 2020) in Praat (Boersma & Weenink, 2022). This approach resulted in continuum steps that sound natural, unlike the highly robotic-sounding stimuli employed in previous distributional learning studies (but see Theodore & Monto, 2019). It also maintained the natural correlations between the most important cues to word-initial stop-voicing in L1-US English (VOT, F0, and vowel duration). Specifically, the F0 at vowel onset of each stimulus was set to respect the linear relation with VOT observed in the original recordings of the talker. The duration of the vowel was set to follow the natural trade-off relation with VOT (Allen & Miller, 1999). Further details on the recording and resynthesis procedure are provided in the SI (§2). A post-experiment survey asked participants: "*Did you notice anything in particular about how the speaker pronounced the different words (e.g. till, dill, etc.)?*" No participant responded that the stimuli sounded unnatural. Analyses reported in the SI (§5.6) found that participants exhibited few attentional lapses ($< 1\%$), including at the start of the experiment ($\leq 1.5\%$). This is a marked improvement over previous studies with robotic sounding stimuli, which elicited high lapse rates, especially at the start of the experiment (12%, Kleinschmidt, 2020). A norming experiment (N = 24 participants) was used to select the three minimal pair continua that differed the least from each other in terms of the categorization responses they elicited (*dill-till*, *din-tin*, and *dip-tip*).

[3] We follow previous work (Kleinschmidt, 2020; Lisker & Abramson, 1964) and refer to pre-voicing as negative VOTs though we note that pre-voicing is perhaps better conceived of as a separate phonetic feature (for discussion, see Mikuteit & Reetz, 2007). This distinction can, for example, be important when interpreting asymmetries in listeners' ability to adapt to left- vs. rightward shifts along the VOT continuum, an issue we revisit in the general discussion.

## 2.3 Procedure

At the start of the experiment, participants acknowledged that they met all requirements and provided consent, as per the Research Subjects Review Board of the University of Rochester. Participants had to pass a headphone test in order to continue (Woods, Siegel, Traer, & McDermott, 2017), and were instructed to not change the volume throughout the experiment. Following instructions, participants completed 234 trials of two-alternative forced-choice categorization. Participants were given the opportunity to take breaks after every 60 trials, which was always during an exposure block. Finally, participants completed an exit survey and an optional demographics survey.

On each of the 234 categorization trials, participants heard a single word spoken by a female talker, and had to click on the word they heard (see Figure 3). Participants were instructed to "answer as quickly and as accurately as possible". Participants were also alerted to the fact that the recordings were subtly different and therefore may sound repetitive. Each trial started with a dark-shaded green fixation dot being displayed. At 500ms from trial onset, two minimal pair words appeared on the screen. At 1000ms from trial onset, the fixation dot would turn bright green and participants had to click on the dot to play the recording. This was meant to reduce trial-to-trial correlations by resetting the mouse pointer to the center of the screen at the start of each trial. Participants responded by clicking on the word they heard and the next trial would begin. Unbeknownst to participants, the 234 trials were split into three exposure blocks (54 trials each) and six test blocks (12 trials each), as shown in Figure 2.

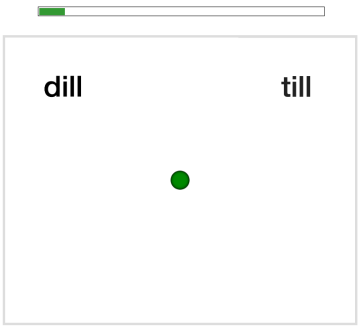


*Figure 3.* Example trial display. When the green button turned bright green, participants had to click on it to play the recording. The placement of response options was counter-balanced across participants.

*Test blocks.* The experiment started with a test block—often assumed, but not tested, to elicit identical response distributions across exposure conditions (see also Colby et al., 2018; Xie, Buxó-Lugo, et al., 2021). Test blocks were identical within and across conditions, always including 12 minimal pair trials assessing participants' categorization at 12 different VOTs (-5, 5, 15, 25, 30, 35, 40, 45, 50, 55, 65, 70 ms). The same brief test block followed each exposure block to assess the effects of cumulative exposure. As alluded to in the introduction, the use of repeated testing introduces procedural challenges.

Three considerations informed the decision to keep testing short. First, listeners' attention span is limited. Second, repeated testing over uniform test continua can reduce or undo the effects of informative exposure (Cummings & Theodore, 2023; Giovannone & Theodore, 2021; Liu & Jaeger, 2018, 2019; Tzeng, Nygaard, & Theodore, 2021; Zheng & Samuel, 2023). Third, the three exposure conditions differ in their exposure distributions, so that the "same" distribution in a test block will convey different information when evaluated relative to these exposure conditions. Theories of error-driven learning and ideal information integration (discussed in the introduction) predict that this affects adaptation. By keeping tests short relative to exposure, we aimed to minimize the influence of test trials on adaptation while still being able to estimate changes in listeners categorization function.

The assignment of VOTs to minimal pair continua was randomized for each participant, but counter-balanced within and across test blocks. Each minimal pair appear equally often within each test block (four times), and each minimal pair appear with each VOT equally often (twice) across all six test blocks (and no more than once per test block). The order of response options—whether the /d/-initial word appeared on the left or right of the screen (see Figure 3)—was held constant within each participant, and counter-balanced across participants.

*Exposure blocks.* Each exposure block consisted of 24 /d/ and 24 /t/ trials, as well as 6 catch trials that served as a check on participant attention throughout the experiment (2 instances for each of three combinations of the three catch recordings). With a total of 144 trials, and intermittent tests after 0, 48, and 96 critical trials, the experiment was designed to measure the effects of exposure at substantially earlier moments than in similar previous experiments (cf. >200 critical trials in Clayards et al., 2008; Kleinschmidt, 2020; Nixon et al., 2016; Theodore

& Monto, 2019).

The distribution of VOTs across the 48 /d/-/t/ trials depended on the exposure condition. We first created a *baseline* condition. We set the VOT means to 5ms for /d/ and 50ms for /t/. We took two steps to increase the ecological validity of the VOT distributions, compared to similar previous work (Clayards et al., 2008; Idemaru & Holt, 2011, 2020; Kleinschmidt, 2020; Kleinschmidt, Raizada, & Jaeger, 2015). First, previous studies exposed each group of listeners to categories with identical variance. We instead set the variance for /d/ to 80 $ms^2$ and for /t/ to 270 $ms^2$ VOT. This qualitatively follows the natural asymmetry in the variance of VOT for /d/ and /t/ found in everyday speech (Chodroff & Wilson, 2017; Docherty, 1992; Lisker & Abramson, 1964).[4] Second, rather than to expose listeners to fully symmetric *designed* distributions that would never be experienced in everyday speech, we *randomly sampled* from the intended VOT distribution (see top row of Figure 4). Specifically, we sampled VOTs for three exposure blocks, and then created three Latin-square designed lists that counter-balanced the order of these blocks across participants.

Following Kleinschmidt et al. (2015), half of the /d/ and half of the /t/ trials in each exposure block were labeled, and the other half were unlabeled. Unlabeled trials were identical to test trials except that the distribution of VOTs across those trials was bimodal (rather than uniform), and determined by the exposure condition (see Figure 4). Labeled trials instead presented two response options with identical stop onsets—e.g., *din* and *dill* to label the input as a /d/. While lexical context often disambiguates and labels sounds in everyday speech (facilitating adaptation, Burchill, 2023; Burchill, Liu, & Jaeger, 2018), disambiguating context is not *always* available. Especially with unfamiliar accents, listeners often have uncertainty about the word sequences they are hearing, reducing the labeling information available to them. Here, we thus struck a compromise between never or always labeling the input.

Next, we created the two additional exposure conditions by shifting all VOTs sampled for the baseline condition by +10 or +40 ms (see Figure 4). This approach exposes participants to

[4] The specific variance values we chose strike a compromise between the variance observed in natural productions (e.g, mean by-talker variances of 29 $ms^2$ for /d/ and 275 $ms^2$ for /t/ in hyper-articulated isolated word productions, and 70 $ms^2$ for /d/ and 410 $ms^2$ for /t/ in connected speech, Chodroff & Wilson, 2017), and the range of natural-sounding VOTs that we were able to generate with our procedure (for VOTs > 130ms, some recordings would not have sounded natural).

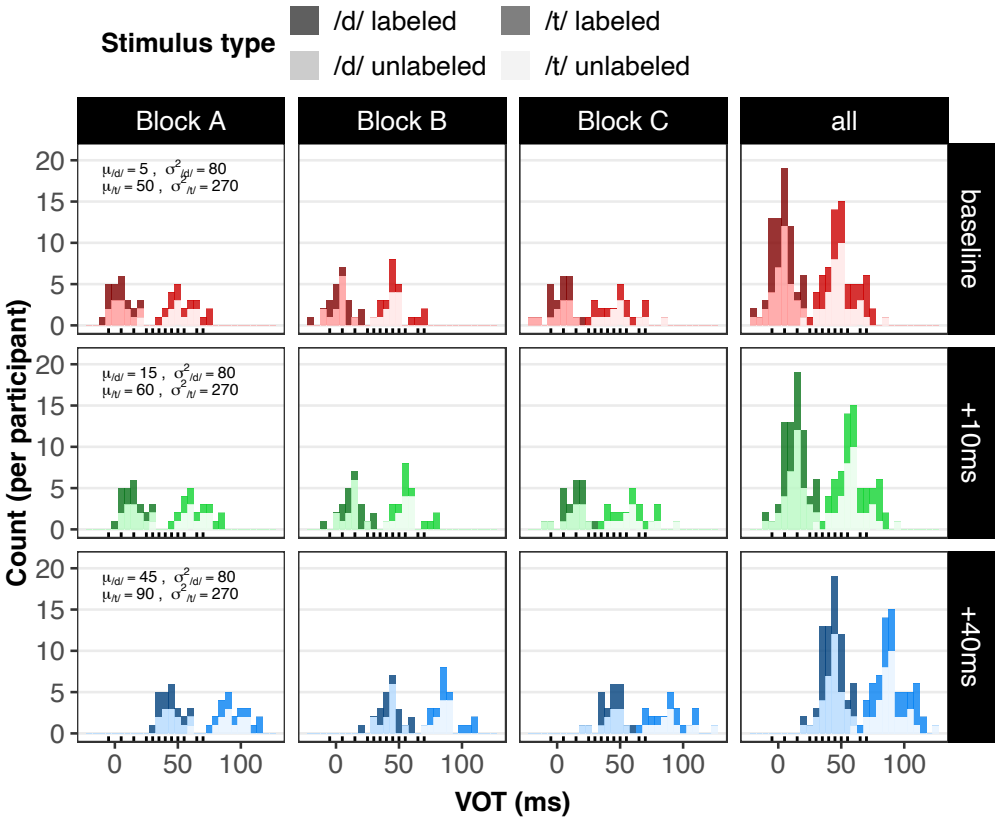


*Figure 4.* Histogram of VOTs for each of the three exposure blocks A-C by exposure condition and trial type (labeled or unlabeled, sampled from /d/ or /t/). Each exposure block contained 12 labeled /d/, 12 labeled /t/, 12 unlabeled /d/, and 12 unlabeled /t/ trials, as well as 6 catch trials (not shown). Except for the shift in VOTs, the VOT distribution of these trials–as well as the relative placement of labeled and unlabeled trials—was identical across exposure conditions. The order of exposure blocks A-C was counter-balanced across participants within each exposure condition using a Latin-square design. Tick marks along the x-axis show the location of the twelve *test* tokens, which were identical across conditions.

heterogeneous *samples* of VOT distributions for /d/ and /t/ that varied across blocks, while holding all aspects of the input constant across conditions except for the shift in VOT—including the placement of labeled and unlabeled trials relative to the exposure condition's category means. The order of trials was randomized within each block and participant, with no more than two catch trials in a row. Participants were randomly assigned to one of 18 lists, crossing 3 (exposure condition) x 3 (block order) x 2 (placement of response options during unlabeled test and exposure trials). We note that the naming of conditions (baseline, +10, +40) should be understood as *relative to each other*, rather than relative to listeners' prior experience.

### 2.4 Exclusions

Exclusion criteria were determined prior to analysis. Due to data transfer errors, four participants' data were not stored and therefore excluded from analysis. We further excluded from analysis participants who committed more than three errors out of the 18 catch trials (<83% accuracy, N = 1), participants who committed more than four errors out of the 72 labelled trials (<94% accuracy, N = 0), participants with an average reaction time more than three standard deviations from the mean of the by-participant means (N = 0), participants who had atypical categorization functions even prior to exposure (N = 2, see SI, §3 for details), and participants who reported not to have used headphones (N = 0). This left for analysis 17,136 exposure and 8,568 test observations from 119 participants (94% of total), approximately evenly split across the three exposure conditions (baseline: 40 participants; +10: 40; +40: 39).

## 3 Results

Below, we begin by describing our analysis approach, which deviates from previous work. We therefore first demonstrate that our approach replicates previous findings when applied to our data at the level of analysis employed in previous work: the detection of overall changes in listeners' categorization behavior after exposure. Following this, we turn to our primary questions: *incremental* changes in participants' categorization responses from pre-exposure onward, depending on the type (exposure condition) and amount of exposure (test block). This allow us to assess predictions (1) and (2a,b) about the role of prior and recent experience in explaining incremental adaptive speech perception, as well as prediction (4) about the diminishing rate of behavioral changes with increasing exposure. To facilitate the interpretation of our results, we introduce normative models (ideal observers) that determine the expected categorization functions of idealized listeners prior to, and following, exposure (following the same approach used in Figure 1). This allows us to identify previously unrecognized constraints on adaptive speech perception (prediction 3 - learning to convergence).

### 3.1 Analysis approach

We analyzed participants' categorization responses during exposure and test blocks in two separate Bayesian mixed-effects psychometric models, using `brms` (Bürkner, 2017) in `R` (R Core Team, 2022; RStudio Team, 2020).[5] Psychometric models account for attentional lapses while estimating participants' categorization functions. Failing to account for attentional lapses—while still common in research on speech perception (for exceptions, see Clayards et al., 2008; Kleinschmidt & Jaeger, 2016)—can lead to biased estimates of categorization boundaries (Prins, 2011; Wichmann & Hill, 2001). For the advantages of Bayesian psychometric models, we refer to Kuss, Jäkel, and Wichmann (2005) and Prins (2019b). For the present experiment, lapse rates were negligible (0.8%, 95%-CI: 0.4 to 1.5%), and all results replicate in simple mixed-effects logistic regressions (T. Florian Jaeger, 2008). This lapse rate compares favorably against those assumed or reported in prior work (e.g., Clayards et al., 2008; Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016).

The psychometric models for exposure and test blocks each regressed participants' categorization responses against the full factorial interaction of VOT, block, and exposure condition, along with the maximal random effect structure (by-participant intercepts and slopes for VOT, block, and their interaction, and by-item intercept and slopes for the full factorial design; see SI, §5.1). All hypothesis tests reported below are based on these models. Figure 5 summarizes the results that we describe in more detail next. Panels A and B show participants' categorization responses during exposure and test blocks, along with the categorization function estimated from those responses via the mixed-effects psychometric models. These panels facilitate comparison between exposure conditions within each block. Panel C summarizes changes in listeners' point of subject equality (PSE)—i.e., the point along the VOT continuum at which participants are equally likely to respond "d" or "t"—across blocks and conditions. This highlights how the type and amount of phonetic input affect listeners' categorization functions. Here we focus on the test blocks, which were identical within and across exposure conditions. Analyses of

[5] For the analyses of test blocks, fitting the models separately removes any potential collinearity between effects of exposure and effects of VOT. The SI reports additional analyses over the combined data, including extensions of the psychometric models to include lapse rates that can vary by block (§5.6) and non-parametric smooths to model non-linear effects of VOT and exposure (§5.9). All analyses replicate the findings reported here.

the exposure blocks, reported in the SI (§5.3), replicate all effects found in the test blocks.

### 3.2 Conceptual replication (averaging over test blocks)

We first use the psychometric mixed-effects model to analyze participants' behavior averaged over all test blocks. This analysis recasts, within a psychometric model, the type of analysis most common in the field: it assesses overall changes in listeners' behavior after exposure, without telling us how these changes accumulate, how they relate to listeners' prior expectations, or how they compare to behavior that would be expected from a learner that has fully adapted to the unfamiliar talker. Prediction (2b) states that changes in listeners' categorization function should depend on the distribution of phonetic cues in the exposure input. Specifically, the +10 condition should elicit a rightward shift in the categorization function relative to the baseline condition, and the +40 condition should elicit an even larger rightward shift. This is also what previous work found (Kleinschmidt & Jaeger, 2016).

Across all test blocks, participants were more likely to respond "t" the longer the VOT ($\hat{\beta} = 15.09$, 90%-CI = $[12.377, 17.625]$, $BF \geq 8000$, $p_{posterior} = 1$). Exposure affected participants' categorization responses in the predicted direction. Marginalizing over Tests 1-6, participants in the +40 condition were less likely to respond "t" than participants in the +10 condition ($\hat{\beta} = -2.26$, 90%-CI = $[-3.258, -1.228]$, $BF = 162.3$, $p_{posterior} = 0.994$) or the baseline condition ($\hat{\beta} = -3.08$, 90%-CI = $[-4.403, -1.669]$, $BF = 215.2$, $p_{posterior} = 0.995$). There was also evidence—albeit less decisive—that participants in the +10 condition were less likely to respond "t" than participants in the baseline condition ($\hat{\beta} = -0.82$, 90%-CI = $[-1.887, 0.282]$, $BF = 8.9$, $p_{posterior} = 0.899$). That is, the +10 and +40 conditions resulted in categorization functions that were shifted rightwards compared to the baseline condition, as also evident in Figure 5A.[6]

Unlike the differences in the relative shift of the categorization function, there was little

[6] The perceptual model contained in the our psychometric mixed-effects model describes the effect of VOT on the log-odds of "t"-responses as a line. The main effect of VOT is the average slope of that line across exposure conditions. The $\hat{\beta}$s for the comparisons across conditions indicate differences in the intercept of that line. Negative $\hat{\beta}$s thus indicate a *down*ward shift of that line in one condition, relative to the other. These downward shifts result in *right*ward shifts of the point of subjective equality (PSE), the VOT at which "t" and "d" responses are equally likely. This also shows in Figure 5A. In this figure, predictions are transformed into proportion "t"-responses and the downward shifts appear visually as a rightward shifts of the S-shaped categorization function (of one condition relative to another).

evidence that the *slopes* of listeners' categorization functions differed between exposure conditions ($0.7 < \text{BFs} < 2.1$; see also Figure 5A). This lack of notable differences in the slope is precisely what is expected under distributional learning models, since our exposure conditions manipulated neither the category variances of /d/ and /t/ nor the distance between their category means. In the remainder of the main text, we thus focus on right vs. left shifts in listeners' categorization function—i.e., changes in listeners' PSE. Parallel analyses of changes (or lack thereof) in the slopes of listeners' categorization functions are reported in the SI (§5.4-§5.5), and do not affect any of our conclusions.

In summary, our initial analysis conceptually replicates previous findings that exposure to different VOT distributions changes listeners' categorization responses (for /b/-/p/: Clayards et al., 2008; Kleinschmidt, 2020; Kleinschmidt et al., 2015; for /g/-/k/, Theodore & Monto, 2019). This replication makes two noteworthy contributions. It extends previous findings to stimuli and stimulus distributions that somewhat more closely resemble those found in everyday speech. And it adds to a small body of work that goes beyond dichotomous comparisons, testing stronger hypotheses about the relative ordering of multiple exposure conditions (e.g., Babel, McAuliffe, Norton, Senior, & Vaughn, 2019; Bejjanki, Clayards, Knill, & Aslin, 2011; Bradlow & Bent, 2008; Cummings & Theodore, 2023; Kleinschmidt, 2020; Liu & Jaeger, 2018).

### 3.3 Overview of incremental analyses

Next, we turn to our primary questions. We assess Incremental changes in participants' categorization responses from three mutually complementing perspectives. First, we compare how exposure affects listeners' categorization responses *relative to other exposure conditions*. This is the perspective taken in previous studies and in the conceptual replication presented above, but extended to test *how early* in the experiment differences between exposure conditions begin to emerge. Second, we compare how exposure *incrementally changes* listeners' categorization responses from block to block within each condition, relative to listeners' responses prior to any exposure. Together with the first perspective, this allows us to test predictions (1)-(2a,b) from the introduction. Third, we compare changes in listeners' responses to those expected from an ideal observer that has fully learned the exposure distributions. This analysis has the potential to

identify constraints on cumulative adaptation. As we show, the latter two perspectives—made possible by the incremental exposure-test paradigm—afford stronger tests of predictions (3 - *learn to convergence*) and (4 - *diminishing returns*), and suggest previously unrecognized constraints on the early moments of adaptive speech perception. For all three analyses, we initially focus on Tests 1-4 with intermittent exposure. Finally, we analyze the effects of repeated testing without intermittent exposure blocks during Tests 4-6. This reveals that such testing has under-appreciated methodological and theoretical consequences.

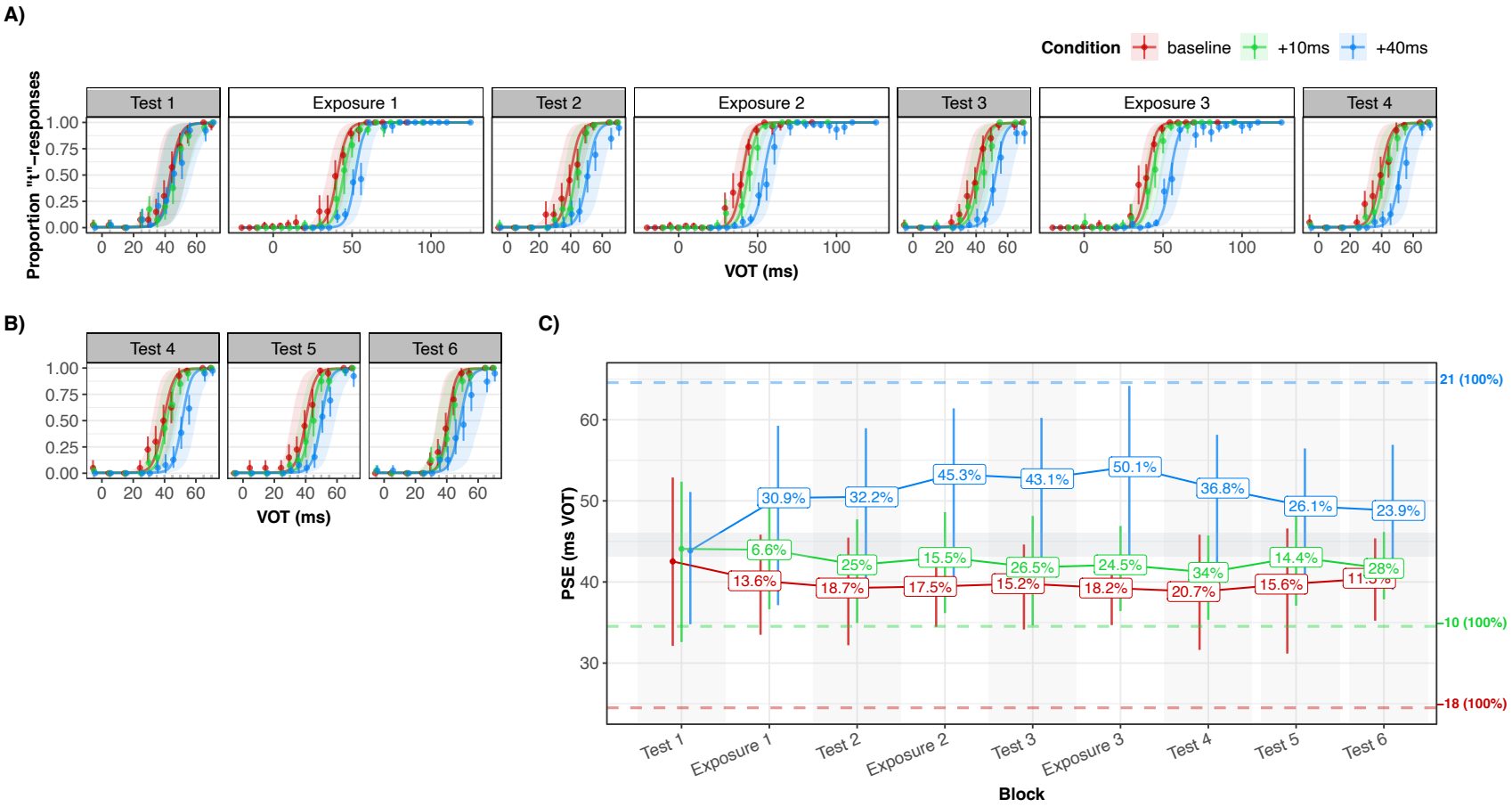


*Figure 5.* Summary of results. **Panel A:** Changes in listeners psychometric categorization functions as a function of exposure, from Test 1 to Test 4 with all intervening exposure blocks (only unlabeled trials were included in the analysis of exposure blocks since labeled trials provide little information about listeners' categorization function). Point ranges indicate the mean proportion of participants' "t"-responses and their 95% bootstrapped CIs. Lines and shaded intervals show the *maximum a posteriori* (MAP) estimates and 95%-CIs of a Bayesian mixed-effects psychometric model fit to participants' responses. **Panel B:** Same as Panel A but for the final three test blocks without intervening exposure. Test 4 is shown as part of both Panels A and B. **Panel C:** Changes across blocks and conditions in listeners' point-of-subjective-equality (PSE) of the lapse-corrected categorization functions from Panels A & B (i.e., the PSE of the perceptual model inferred from listeners' responses; for changes in the slope of that function, see SI, §4.3). Point ranges represent the posterior medians and their 95%-CIs derived from the psychometric model. Horizontal dashed lines indicate 95%-CIs of the PSEs expected from an idealized learner (an ideal observer model that has fully learned the exposure distributions). Percentage labels indicate the degree of shift in PSE exhibited by participants as a proportion of the expected shift under the idealized learners (for details, see SI, §5.8.2). Horizontal gray ribbon indicates the 95%-CIs of the PSEs expected from an idealized listener *prior to any exposure*.

### 3.4 How quickly does exposure affect listeners' categorization responses? (comparing exposure conditions within each test block)

Table 2
*The simple effects of the exposure conditions for each test block. This analysis asks how early exposure starts to affect participants' categorization responses, and when (if ever) these changes were undone with repeated testing. Note that rightward shifts of the categorization function (and its PSE) correspond to negative estimates (lower intercepts in predicting the log-odds of "t"-responses). Predicted nulls for Test 1 were tested using the Savage-Dickey density ratio.*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Test block 1 (pre-exposure)** | | | | | |
| $PSE_{+10} = PSE_{baseline}$ | -0.34 | 0.75 | [-2.03, 1.44] | 3.3 | 0.77 |
| $PSE_{+40} = PSE_{+10}$ | 0.25 | 0.73 | [-1.34, 1.9] | 3.7 | 0.79 |
| $PSE_{+40} = PSE_{baseline}$ | -0.08 | 0.91 | [-2.12, 2.08] | 4.8 | 0.83 |
| **Test block 2** | | | | | |
| $PSE_{+10} > PSE_{baseline}$ | -1.45 | 0.89 | [-2.93, 0.18] | 13.7 | 0.93 |
| $PSE_{+40} > PSE_{+10}$ | -2.08 | 0.99 | [-3.82, -0.17] | 24.3 | 0.96 |
| $PSE_{+40} > PSE_{baseline}$ | -3.49 | 1.24 | [-5.63, -1.07] | 54.2 | 0.98 |
| **Test block 3** | | | | | |
| $PSE_{+10} > PSE_{baseline}$ | -0.78 | 0.62 | [-1.89, 0.36] | 7.9 | 0.89 |
| $PSE_{+40} > PSE_{+10}$ | -2.80 | 0.82 | [-4.19, -1.11] | 86.0 | 0.99 |
| $PSE_{+40} > PSE_{baseline}$ | -3.56 | 0.97 | [-5.2, -1.58] | 110.1 | 0.99 |
| **Test block 4** | | | | | |
| $PSE_{+10} > PSE_{baseline}$ | -0.88 | 0.85 | [-2.36, 0.85] | 4.8 | 0.83 |
| $PSE_{+40} > PSE_{+10}$ | -3.32 | 0.89 | [-4.88, -1.64] | 128.0 | 0.99 |
| $PSE_{+40} > PSE_{baseline}$ | -4.16 | 1.21 | [-6.28, -1.88] | 122.1 | 0.99 |
| **Test block 5 (repeated testing without additional exposure)** | | | | | |
| $PSE_{+10} > PSE_{baseline}$ | -1.33 | 0.71 | [-2.56, 0] | 19.2 | 0.95 |
| $PSE_{+40} > PSE_{+10}$ | -2.39 | 0.86 | [-3.89, -0.8] | 65.1 | 0.98 |
| $PSE_{+40} > PSE_{baseline}$ | -3.72 | 1.01 | [-5.5, -1.7] | 139.4 | 0.99 |
| **Test block 6 (repeated testing without additional exposure)** | | | | | |
| $PSE_{+10} > PSE_{baseline}$ | -0.22 | 0.72 | [-1.48, 1.11] | 1.7 | 0.62 |
| $PSE_{+40} > PSE_{+10}$ | -1.70 | 0.79 | [-3.08, -0.17] | 25.0 | 0.96 |
| $PSE_{+40} > PSE_{baseline}$ | -1.91 | 0.99 | [-3.63, 0.02] | 18.5 | 0.95 |

Figure 5A suggests that differences between exposure conditions emerged very early in the experiment. This is confirmed by Bayesian hypothesis tests summarized in Table 2, which we discuss next. Prior to any exposure, during Test 1, participants' responses did not differ across exposure conditions. This result is predicted by models of adaptive speech perception under the assumptions that (a) participants in the different groups have similar prior experiences, and that

(b) our sample size is sufficiently large to yield stable estimates of listeners' categorization function.[7] Equality of pre-exposure behavior across exposure groups is also implicitly assumed—but rarely tested—in the interpretation of most studies on adaptive speech perception (when it is tested, it often turns out that this assumption is *not* necessarily warranted, presumably due to insufficient sample sizes, cf. Kleinschmidt, 2020).

During Test 2, after exposure to only 24 /d/ and 24 /t/ stimuli (thereof half labeled), participants' categorization responses already differed between exposure conditions (BFs $> 14$). All differences between exposure conditions that emerged at Test 2 followed prediction (2b - *exposure distributions*). Additional analyses reported in the SI (§5.3) found that listeners' categorization functions had already changed in the predicted direction during the first *exposure* block, in line with Figure 5A. This suggests that changes in listeners' categorization responses emerged quickly at the earliest point tested—after only a fraction of exposure trials previously tested in similar paradigms.

The effects of the three exposure conditions persisted until Test 4, always in line with prediction (2b). Table 2 does, however, indicate an interesting non-monotonic development. While the difference between the +40 condition and both the baseline and +10 condition continued to increase numerically with increasing exposure (increasingly larger magnitude of negative estimates in Tests 2-4), the same was not the case for the difference between the +10 and the baseline condition. Instead, the difference between the +10 and baseline condition *reduced* with increasing exposure (while maintaining its direction; from -1.4 to -0.83, see Table 2). At first blush, this non-monotonicity appears to contradict prediction (2a) that the magnitude of exposure effects should increase with increasing exposure. In the next section, we show that the results do, in fact, *support* prediction (2a) when listeners' prior expectations are considered. Indeed, the seemingly unexpected non-monotonicity—which would be impossible to detect without repeated testing—turns out to be important for understanding incremental adaptation.

[7] The moderate BFs for these hypothesis tests are due to our use of regularizing priors, which have non-negligible density over the null (for an introduction to the Savage-Dickey method, see Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). Rather than to test the null against more plausible alternative priors, which would predictably increase the evidence for the null, we appeal to readers' intuition: the 90%-CIs of the comparisons for Test 1 are all approximately centered around zero; there is very little evidence in favor of an effect in either direction.

### 3.5 Incremental adaptation from prior expectations (comparing block-to-block changes within exposure conditions)

Next, we compare how listeners' categorization responses changed from block to block *within* each exposure condition. This allows us to understand changes in listeners' categorization function relative to listeners' pre-exposure behavior, thereby assessing the joint effects of predictions (1 - *prior expectations*) and (2a,b - *exposure amount & distributions*). To facilitate visual comparison across blocks and conditions, Figure 5C summarizes the block-to-block changes in listeners' PSE. Focusing for now on Tests 1-4, this highlights three aspects of participants' behavior that were not readily apparent in the statistical comparisons presented so far.

Table 3
*Was there incremental change from test block 1 to 4? Did these changes dissipate with repeated testing from block 4 to 6? This table summarizes the simple effects of block for each exposure condition. Note that rightward shifts of the categorization function (and its PSE) correspond to negative estimates (lower intercepts in predicting the log-odds of "t"-responses).*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Difference between blocks: baseline** | | | | | |
| Block 1 to 2: decreased PSE | 1.16 | 0.71 | [-0.22, 2.52] | 12.9 | 0.93 |
| Block 2 to 3: decreased PSE | 0.12 | 0.70 | [-1.31, 1.48] | 1.3 | 0.57 |
| Block 3 to 4: decreased PSE | 0.17 | 0.53 | [-0.86, 1.12] | 1.7 | 0.63 |
| *Block 1 to 4: decreased PSE* | 1.48 | 1.13 | [-0.73, 3.44] | 7.6 | 0.88 |
| Block 4 to 5: increased PSE | -0.37 | 0.49 | [-1.28, 0.53] | 3.5 | 0.78 |
| Block 5 to 6: increased PSE | -0.57 | 0.61 | [-1.65, 0.62] | 4.6 | 0.82 |
| *Block 4 to 6: increased PSE* | -0.94 | 0.73 | [-2.29, 0.51] | 7.2 | 0.88 |
| **Difference between blocks: +10** | | | | | |
| Block 1 to 2: decreased PSE | 0.16 | 0.79 | [-1.17, 1.62] | 1.4 | 0.59 |
| Block 2 to 3: decreased PSE | 0.60 | 0.66 | [-0.57, 1.85] | 4.5 | 0.82 |
| Block 3 to 4: decreased PSE | 0.17 | 0.77 | [-1.32, 1.64] | 1.4 | 0.58 |
| *Block 1 to 4: decreased PSE* | 0.94 | 1.21 | [-1.3, 3.17] | 3.5 | 0.78 |
| Block 4 to 5: increased PSE | -0.58 | 0.58 | [-1.63, 0.52] | 4.9 | 0.83 |
| Block 5 to 6: increased PSE | 0.44 | 0.65 | [-0.79, 1.65] | 0.3 | 0.24 |
| *Block 4 to 6: increased PSE* | -0.12 | 0.83 | [-1.63, 1.48] | 1.3 | 0.56 |
| **Difference between blocks: +40** | | | | | |
| Block 1 to 2: increased PSE | -2.06 | 0.79 | [-3.43, -0.56] | 45.2 | 0.98 |
| Block 2 to 3: increased PSE | -0.73 | 0.78 | [-2.09, 0.63] | 4.7 | 0.83 |
| Block 3 to 4: increased PSE | -0.06 | 0.81 | [-1.48, 1.34] | 1.1 | 0.53 |
| *Block 1 to 4: increased PSE* | -2.86 | 1.12 | [-4.87, -0.73] | 50.3 | 0.98 |
| Block 4 to 5: decreased PSE | 0.61 | 0.77 | [-0.75, 1.93] | 3.6 | 0.78 |
| Block 5 to 6: decreased PSE | 0.75 | 0.72 | [-0.56, 2] | 5.5 | 0.85 |
| *Block 4 to 6: decreased PSE* | 1.36 | 0.95 | [-0.33, 2.99] | 10.3 | 0.91 |

First, while the PSEs for the +40 and +10 conditions were shifted rightwards compared to the baseline condition, both the +10 and the baseline condition seem to shift *left*wards relative to their pre-exposure starting point in Test 1. Bayesian hypothesis tests summarized in Table 3 find moderate support for a leftward shift from Test 1 to 4 in both the +10 condition (3.5) and the baseline condition (7.6). In contrast, there was strong support that the +40 condition shifted rightwards relative to pre-exposure (0). To understand this pattern, it is helpful to relate the three exposure conditions to the phonetic distribution in listeners' prior experience. Figure 6 shows the exposure means for /d/ and /t/ relative to the distributions of three important cues to the word-initial /d/-/t/ contrast in L1-US English (based on databases of isolated and connected speech, Chodroff & Wilson, 2018). This comparison offers an explanation as to why the baseline condition (and to some extent the +10 condition) shift leftwards with increasing exposure, whereas the +40 condition shifts rightwards: relative to the distribution of VOT for /d/ and /t/ in listeners' prior experience, only the +40 condition presents category means that are clearly larger than expected along VOT, whereas the baseline condition and, to some extent, the +10 condition presented lower-than-expected category means. That is, once we take into account how our exposure conditions relate to listeners' prior experience (prediction 1), both the direction of changes from Test 1 to 4 *within* each exposure condition (Table 3), and the direction of differences *between* exposure conditions receive an explanation (Table 2). To further illustrate this point, the horizontal gray ribbon in 5C shows the range of PSEs predicted by Bayesian ideal observers trained on the distribution of VOT, f0, and vowel duration for isolated word productions in Figure 6 (for details, see SI, §4.1).

Second, we find support for prediction (4) about the *diminishing returns* of additional exposure predicted by some theories of adaptive speech perception. The estimates in Table 3 suggest that listeners' PSEs changed most substantially from Test 1 to Test 2, and then changed less and less with additional exposure up to Test 4 (smaller magnitude of estimates compared to earlier test blocks). This seems to be particularly pronounced for the baseline condition and the +40 condition—the two conditions that exhibited the largest shifts relative to pre-exposure. As mentioned in our Open Science statement, our experiment was not designed to have high power to assess such *changes in the magnitude of the shifts* across the block within each condition. We

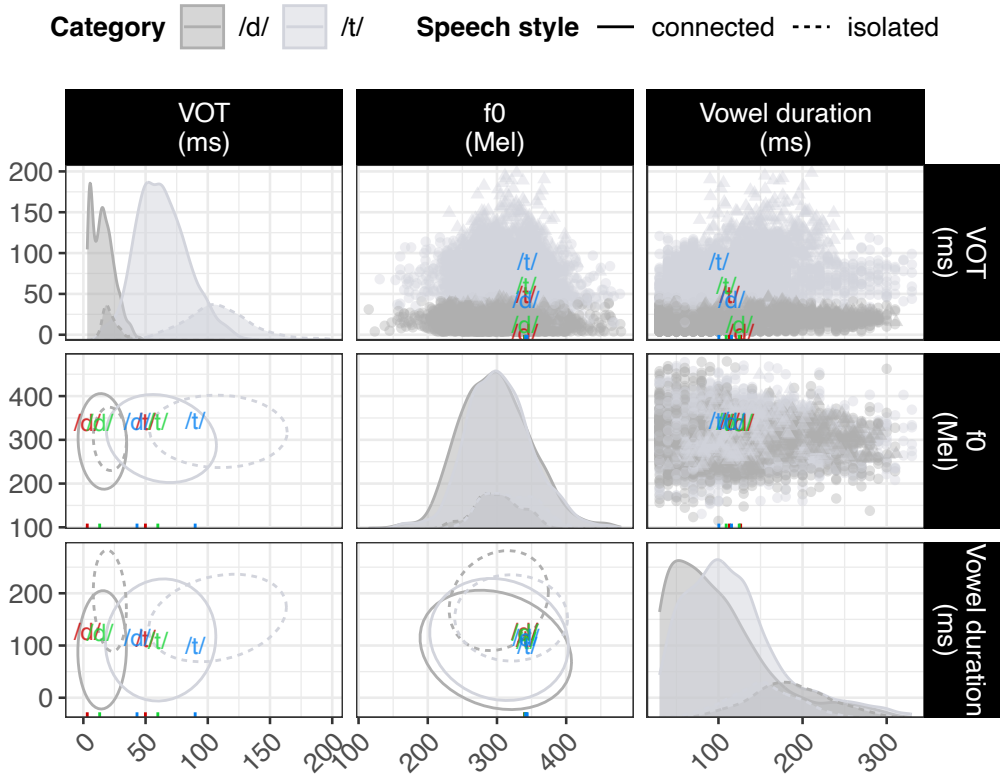


*Figure 6.* Placement of exposure stimuli relative to an estimate of typical phonetic distributions for 5,756 word-initial /d/ and /t/ productions by 92 female L1 talkers of US English in Chodroff and Wilson (2018). After voice onset time, f0 and vowel duration are two of the most informative cues to word-initial /d/-/t/ in L1 US English. For details, see SI §4.1. Colored labels show the category means of the exposure conditions.

did, however, conduct post-hoc hypothesis tests to assess the support for this pattern. These tests found anecdotal to moderately strong evidence in support of prediction (4 - *diminishing returns*). For the +40 condition, the shift from Test 1 to 2 was larger than the shift from Test 2 to 3 ($\hat{\beta} = -1.31$, 90%-CI = $[-3.518, 0.904]$, $BF = 5.4$, $p_{posterior} = 0.845$), which was larger than the shift from Test 3 to 4 ($\hat{\beta} = -0.67$, 90%-CI = $[-2.861, 1.546]$, $BF = 2.4$, $p_{posterior} = 0.702$). Comparing the change from Test 1 to 2 against the change from Test 3 to 4, there was stronger support that the speed of changes in the PSE decreased ($\hat{\beta} = -1.98$, 90%-CI = $[-4.067, 0.2]$, $BF = 14.9$, $p_{posterior} = 0.937$). For the baseline condition, the shift from Test 1 to 2 was larger than the shift from Test 2 to 3 ($\hat{\beta} = 1.05$, 90%-CI = $[-1.046, 3.142]$, $BF = 4.4$, $p_{posterior} = 0.814$), which was almost identical, but slightly smaller, than the shift from Test 3 to 4 ($\hat{\beta} = 0.04$, 90%-CI = $[-1.805, 1.86]$, $BF = 1.1$, $p_{posterior} = 0.515$). Again, a comparison of the change from Test 1 to

2 against the change from Test 3 to 4, yielded the strongest support that the speed of changes in the PSE decreased ($\hat{\beta} = 1.01$, 90%-CI = $[-0.745, 2.821]$, $BF = 5.3$, $p_{posterior} = 0.842$). For both the +40 and the baseline condition, there was only anecdotal evidence that the final exposure block resulted in *any additional* shift in listeners' PSE (BFs $\leq 1.7$, cf. Table 3).

Third and finally, Panel C also begins to illuminate the reasons for the non-monotonic development of the +10 and baseline conditions relative to each other, discussed in the previous section. In particular, this non-monotonicity does *not* appear due to a reversal of the effects in either of the two exposure conditions. Rather, both exposure conditions continue to change listeners' categorization function in the same direction from Test 1 to Test 4, in line with predictions (2a) and (2b). However, after the rapid change from the pre-exposure Test 1 to the first post-exposure Test 2, listeners' categorization responses in the baseline condition did not change as much as in the +10 condition. In fact, listeners' categorization function in the baseline condition seems to have plateaued after the first exposure block.

This explains the reduction in the difference between the +10 and baseline conditions discussed in the previous section. It does, however, raise the question *why* listeners' responses in the baseline condition did not change further with increasing exposure. One explanation would be that participants in the baseline condition did for some reason—including chance—fully learn the relevant phonetic distributions within a single block of exposure, whereas participants in the +10 condition were not. The third and final perspective we provide on incremental changes in participants' behavior suggests that this was *not* the case.

### 3.6 Constraints on cumulative adaptation (comparing exposure effects against idealized learner models)

Beyond the perspectives on incremental adaptation discussed so far, Figure 5C compares participants' responses against those of an idealized learner that has fully learned the exposure distributions (colored dashed lines). Specifically, we fit Bayesian ideal observers against the labeled VOT distributions of each exposure condition, using the same approach used for the idealized pre-exposure listeners (horizontal gray ribbons). The dashed lines show the PSEs expected from such idealized learners (for details, see SI §4.3). This approach follows previous

Table 4
*Did participants* not *converge against the PSE expected from idealized learner? This table compares changes in participants' categorization function against those expected from idealized learners. This table summarizes results for the test block following the final exposure block (Test 4). For identical tests for all test blocks, see SI (§5.8.1).*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| $\lvert\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{baseline}})\rvert > 0$ | 0.32 | 0.06 | [0.2, 0.44] | $\geq 8000$ | 1 |
| $\lvert\Delta(PSE_{ideal_{+10}}, PSE_{actual_{+10}})\rvert > 0$ | 0.15 | 0.04 | [0.05, 0.23] | $\geq 8000$ | 1 |
| $\lvert\Delta(PSE_{ideal_{+40}}, PSE_{actual_{+40}})\rvert > 0$ | 0.29 | 0.05 | [0.2, 0.46] | $\geq 8000$ | 1 |

work (Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016), and makes it possible to assess how far listeners have converged against the exposure distributions. The relevant hypothesis tests are summarized in Table 4.

Figure 5C suggests that listeners in all three exposure conditions did *not* fully learn the exposure distributions ($|\Delta(PSE_{ideal}, PSE_{actual})| > 0$: all BFs $\geq 8000$). By itself, a failure to converge against the performance of an idealized learner would not necessarily constitute evidence against prediction (3 - *learn to convergence*). Listeners might simply not have received sufficient exposure to have learned the exposure distribution. However, as already described for the baseline condition, participants' behavior changed little, if at all, after the first exposure block. Instead, participants seem to have prematurely converged against stable behavior long before they had fully learned the exposure distribution.

The percentage labels in Figure 5C quantify the degree to which participants adapted their PSE towards the statistics of the exposure condition: 0% would correspond to no change relative to the listeners' PSE in Test 1, and 100% would correspond to the PSE predicted for an idealized learner who has fully converged against the exposure distributions. In the baseline condition, changes in participants' PSE seem to converge against approximately 20.7% of what is expected from an idealized learner. A similar pattern of premature convergence is evident for the +40 condition: changes in participants' PSEs seem to have leveled off by Test 4, despite the fact that participants' PSEs had shifted only about half way to the idealized learner's PSE. (For the +10 condition, it is less clear whether participants had already converged against a PSE.) That is, in terms of the possible adaptation scenarios depicted in Figure 1 in the introduction, it seems that

our results most closely resemble the scenario shown in the right-most column of panel B.[8]

Of note, *premature convergence* negatively affected participants' recognition accuracy: while incremental adaptation substantially improved participants' recognition accuracy compared to their pre-exposure accuracy, only participants in the +10 condition came close to achieving the theoretical upper bound expected of an idealized learner. Listeners in the baseline and +40 condition appear to have stopped adapting even though further adaptation would have improved their recognition accuracy.[9]

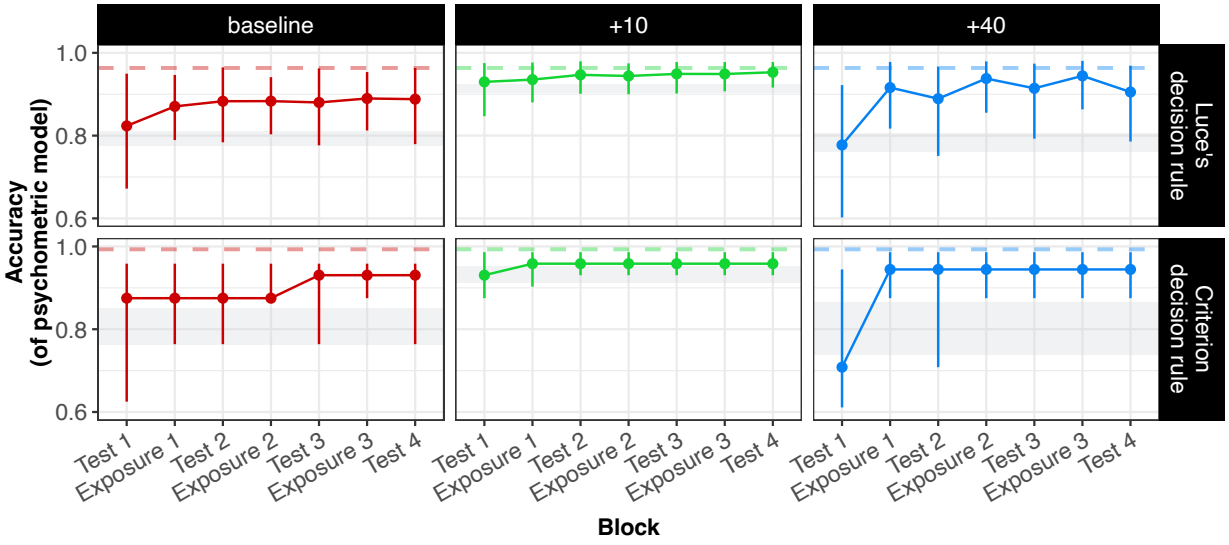


*Figure 7.* Changes across blocks and conditions in participants' recognition accuracy for the unfamiliar talker's speech. For each block, we used participants' categorization functions—estimated by the psychometric mixed-effects model fit to participants' responses—to categorize all 144 exposure inputs of that exposure condition. Accuracy was calculated for the two decision rules that are most commonly assumed to underlie speech recognition as well as perceptual decision-making in other domains (for review, see Massaro & Friedman, 1990): Luce's choice rule (responding proportional to posterior probability of category) and the criterion choice rule (always responding with the category that has highest posterior probability). As in Figure 5C, point ranges represent the posterior medians and their 95%-CIs derived from the psychometric model. Horizontal dashed lines indicate accuracy expected from an idealized learner (an ideal observer model that has fully learned the exposure distributions) and horizontal shaded ribbons indicate the 95%-CI expected from an idealized pre-exposure listener.

[8] Figure 5C would also seem to suggest that the degree of convergence differed between exposure conditions. Two previous studies have observed similar differences, with more extreme exposure shifts eliciting *proportionally* smaller changes in PSEs than less extreme exposure shifts (Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016). For the present data, we found only anecdotal support for this pattern (see SI, §5.8.2).

[9] While a failure to improve from, say, 90% and 95% accuracy might not seem noteworthy, it implies misunderstanding one in ten vs. one in twenty words, thus *doubling* the odds of recognition errors.

If confirmed, premature convergence against stable behavior despite only partial adaptation would challenge prediction (3) of existing distributional learning models. Premature convergence is also unexpected under any other model of adaptive speech perception. In the general discussion, we present an extension to distributional learning models that explains premature convergence, and highlights a striking link between adaptive changes in speech perception and second language learning. As part of that discussion, we present additional models, and entertain methodological artifacts, and analysis confounds that would offer alternative explanation of premature convergence.

### 3.7 Effects of repeated testing

The final hypotheses tests we present investigate the effects of repeated testing. Distributional learning models predict that test stimuli, too, can form part of the input that listeners adapt to. To the extent that the information provided by test stimuli differs from that provided by exposure stimuli, these theories thus predict that repeated testing affects listeners' behavior. And, in a design like ours, with identical test stimuli across conditions, the effects of repeated testing are predicted to differ across conditions. Specifically, with sufficient repetition all conditions would be expected to converge against the distribution of the test stimuli, and thus towards identical behavior across conditions. Theories of error-based learning and ideal information integration further predict that the speed with which repeated testing changes listeners' behavior depends on the degree to which the distribution of test stimuli differs from the distribution of exposure stimuli (Davis & Sohoglu, 2020; Kleinschmidt & Jaeger, 2015; for relevant discussion, see also Lancia & Winter, 2013).

In line with these theories, Figure 5C shows that the effects of exposure reduced from Test 4 to Test 6, and did so primarily for the exposure conditions that differed most from the distribution of test stimuli. In Table 3, this is evident in a reversal of the direction of the block-to-block changes for Tests 5-6, compared to Tests 1-4. For the +40 exposure condition, these block-to-block changes went from rightward shifts in Tests 1-4 to leftward shifts in Tests 5-6 (BF = 10). For the baseline condition, block-to-block changes went from leftward to rightward shifts (BF = 7.2 ). The only exposure condition for which no clear reversal was observed is the

+10 condition (BF = 7.9). As would be expected under theories of error-based learning and ideal information integration, the marginal distribution of VOT during test blocks (mean = 35.8 ms, SD = 22.2 ms) most closely resembled the exposure distribution of the +10 condition (mean = 36.5, SD = 25.9), compared to the baseline (mean = 26.5 ms, SD = 25.9) or +40 condition (mean = 66.5 ms, SD = 25.9).

The effects of repeated testing replicate previous findings from lexically-guided perceptual learning paradigms (Cummings & Theodore, 2023; Giovannone & Theodore, 2021; Liu & Jaeger, 2018, 2019; Reinisch & Holt, 2014; Scharenborg & Janse, 2013; Zheng & Samuel, 2023), and extends them to distributional learning paradigms (see also Colby et al., 2018; Kleinschmidt, 2020). Indeed, the effects of repeated testing can be substantial: while the effects of the +40 condition relative to the other two exposure conditions were reduced but still credible even in Test 6 (BFs $>$ 18), this was no longer the case for the effect of the +10 condition relative to the baseline condition (BF = 1.7; see Table 2). One important methodological implication for future work is that longer test phases do not necessarily increase the statistical power to detect effects of adaptation (unless analyses account for the effects of repeated testing, as done in, e.g., Liu & Jaeger, 2018). Analyses that average over all test tokens—as is still the norm—are bound to systematically underestimate the true adaptivity of human speech perception.

# 4 General discussion

Over the last 20+ years, landmark studies in adaptive speech perception have demonstrated that listeners' interpretation of speech is not static. Instead, it can change with recent exposure, accommodating differences in pronunciation across talkers (for reviews, see Bent & Baese-Berk, 2021; Schertz & Clare, 2020). Research on accent adaptation (AA, Eisner, Melinger, & Weber, 2013; Schertz et al., 2016; Xie et al., 2017), perceptual learning (VGPL/LGPL, Eisner & McQueen, 2005; Kraljic & Samuel, 2006; Kurumada, Brown, & Tanenhaus, 2018; Norris, McQueen, & Cutler, 2003; Reinisch & Holt, 2014; Vroomen et al., 2007), and distributional learning (DL, Bejjanki et al., 2011; Idemaru & Holt, 2020; Kleinschmidt, 2020; Nixon et al., 2016; Theodore & Monto, 2019) suggests that this flexibility is achieved through changes in listeners'

*categorization functions*—the mapping from acoustic or phonetic cues to the phonological categories that form the input to spoken language understanding: after exposure to an unfamiliar talker, listeners interpret physically identical speech input differently.

Here, we have responded to recent calls to better characterize *how* these changes in listeners' categorization functions come about. We set out to test several basic predictions of distributional learning models. Distributional learning models implement the hypothesis that listeners learn the statistics of talkers' speech, and use this implicit knowledge to interpret subsequent utterances by the same talker (Clayards et al., 2008; Idemaru & Holt, 2011; Kleinschmidt & Jaeger, 2015; McMurray & Jongman, 2011). To this end, we modified a distributional learning paradigm to test four predictions about the incremental unfolding of adaptation.

We found that listeners' categorization functions changed incrementally with exposure. The direction and magnitude of that change depended on listeners' prior expectations based on previously experienced speech input from other talkers (prediction 1 - *prior expectations*), and both the amount and distribution of phonetic evidence in the exposure input from the unfamiliar talker (predictions 2a and 2b - *exposure amount & distributions*, respectively). The Bayesian hypothesis tests we conducted also suggest properties of adaptive speech perception that go beyond these qualitative predictions. These properties further inform theory by characterizing the computational properties of the mechanisms underlying listeners' adaptivity. First, we found that participants' categorization functions changed quickly with exposure, and that the speed of these changes slowed-down with additional exposure (prediction 4 - *diminishing returns*). Second, we found evidence of potential constraints on listeners' adaptivity, as well as asymmetries in these constraints depending on whether exposure distributions were shifted downwards or upwards on the VOT continuum—contrary to prediction (3 - *learn to convergence*).

Next, we discuss these findings in more detail, and consider their implications for theories of adaptive speech perception. We begin with predictions (1) and (2a,b). Then we turn to questions about the rate of changes (prediction 4), before considering potential constraints on the initial moments of adaptive speech perception (prediction 3). We close by considering limitations of the present study and how future work can overcome them.

## 4.1 Incremental adaptation based on the amount and distribution of phonetic evidence (Predictions 1 and 2a,b)

To the best of our knowledge, the present study is the first to assess how the joint effects of prior and recent exposure gradiently unfold with increasing exposure, testing predictions (1) and (2a,b). While most contemporary theories of adaptive speech perception share these qualitative predictions, few experiments have investigated how listeners' categorization functions change with continued exposure to a phonetic distribution. Next, we discuss notable exceptions to this trend, and how our results relate to those works.

### 4.1.1 Prediction 1: Adaptation begins with, and integrates, listeners' prior experience

The inclusion of a pre-exposure test in our design made it possible to assess prediction (1)—that shifts in listeners' categorization function should depend on how the exposure distributions *relative to listeners' relevant prior experiences*. This prediction received support by the fact that listeners' responses prior to exposure were well approximated based on a database of /d/ and /t/ productions. While these effects of prior knowledge are often assumed, the present experiment is—to our knowledge—the first time they have been demonstrated for adaptive speech perception.

We emphasize that future tests are necessary to confirm the effects of prior experience observed in the present study. We employed an estimate of participants' prior expectations that is based on a phonetic database of syllable-initial stop productions by speakers of L1-US English. While we chose this database because it matches the speech style of our stimuli, and contained talkers of the same gender (and of relatively similar f0) as the talker used in our experiment, the database is comparatively small. We aimed to remedy this downside by using five-fold cross-validation. This allowed us to express uncertainty about the true range of a typical L1-US English listeners' prior expectations (the gray ribbons in Figure 5C). It does not, however, remove the need to validate our results based on new participants and larger phonetic databases that are more likely to reflect the prior experience of a 'typical' listener.

### 4.1.2 Prediction 2a: Adaptation increases with the amount of exposure

Our results also support prediction (2a)—that the magnitude of changes in listeners' categorization function should gradiently increase with the *amount* of exposure. This prediction received support from the comparisons across blocks: increasing exposure consistently yielded additional shifts in listeners' PSE. This replicates in a DL paradigm recent findings from VGPL/LGPL experiments (Cummings & Theodore, 2023; Kleinschmidt & Jaeger, 2012; Liu & Jaeger, 2018, 2019; Vroomen et al., 2007). As discussed in the introduction, DL paradigms differ from traditional VGPL/LGPL designs in the degree of control they provide over the phonetic distributions in the exposure input. In particular, the Latin-square design over exposure blocks that we used in the present study completely de-correlates the amount of exposure that participants in a given exposure condition received from the specific distribution of phonetic properties (cf. Figure 4). This contrasts with existing LGPL studies, in which exposure amount was confounded with differences in exposure distributions (Cummings & Theodore, 2023; Liu & Jaeger, 2018, 2019), and VGPL studies in which exposure was limited to many repetitions of a single stimulus (e.g., Kleinschmidt & Jaeger, 2012; Vroomen et al., 2007).

In VGPL/LGPL paradigms, listeners are exposed to natural recordings of one phonetic category (e.g., /s/) and shifted instances of a second category that are manipulated to be perceptually more similar to the first category (e.g., /s/-like /ʃ/). Both the typical and the shifted sound instances are lexically or visually labeled by their context. For example, in an LGPL study, the lexical context will disambiguate the intended category of both the typical sounds (e.g., "dino*s*aur") and the shifted sounds (e.g., "medi*sh*ine"). Studies like this typically compare two groups of listeners that differ only in which of the two sounds was shifted. For example, one group of listeners might be exposed to 20 typical /s/ and 20 shifted /s/-like /ʃ/, mixed with 160 filler words that do not contain either sound. The other group of listeners might be exposed to 20 typical /ʃ/ and 20 shifted /ʃ/-like /s/, mixed with the same 160 filler words. Following exposure, the two groups of listeners will categorize sounds along an unlabeled test continuum (e.g., "asi" to "ashi") differently. Specifically, listeners in each group will categorize more sounds along the continuum as belonging to the category that was shifted during exposure (Eisner & McQueen, 2005; Kraljic & Samuel, 2005; e.g., Norris et al., 2003).

In a particularly informative study, Cummings and Theodore (2023) compared shifts in categorization function between groups of listeners after exposure to 1, 4, 10, or 20 lexically labeled shifted /s/ or /ʃ/ tokens (each matched by an equal number of unshifted tokens from the opposite category). Shifts in listeners' categorization functions increased with the number of exposure to tokens, in line with prediction (2a) of distributional learning models. Vroomen et al. (2007) found similarly increasing shifts in categorization functions *within* participants, comparing the effects of 1, 2, 4, …, 32 exposures to visually labeled shifted tokens (see also Kleinschmidt & Jaeger, 2012).[10] The present study demonstrated similarly gradient effects with increasing exposure to a mixture of labeled and unlabeled exposure tokens that were randomly sampled from a *distribution* of phonetic tokens, more closely resembling the situation listeners would experience in everyday speech perception. This aspect of our results thus adds to a growing number of similarities in the findings between LGPL/VGPL and DL paradigms, as expected under the hypothesis that changes observed in both paradigms originate in the same underlying mechanisms (see discussions in Kleinschmidt et al., 2015; Xie et al., 2023; Zheng & Samuel, 2020).

#### 4.1.3 Prediction 2b: Adaptation depends on the phonetic distribution in the exposure input

Prediction (2b)—that the direction and magnitude of changes in listeners' categorization function should depend on the *phonetic distribution* of the exposure input—is perhaps the best documented prediction of the ones we tested (Bejjanki et al., 2011; Chládková et al., 2017; Clayards et al., 2008; Colby et al., 2018; Kleinschmidt & Jaeger, 2011, 2012, 2016; Nixon et al., 2016; Saltzman & Myers, 2021; Theodore & Monto, 2019). For example, Kleinschmidt and Jaeger (2016) exposed five different groups of listeners to VOT distributions for /b/ and /p/ that were shifted to different degrees. The five different exposure conditions were each shifted by 10ms in VOT relative to the other, but held constant the distance between the /b/ and /p/ mean (always 40ms) and the variance of /b/ and /p/ (both always 8.3ms$^2$). All groups of listeners were exposed

[10] With increasing exposure, the direction of shift begins to reverse (returning to baseline after 128-256 exposures, Kleinschmidt & Jaeger, 2011; Vroomen et al., 2007) and can even change directional altogether, depending on the degree of the shift (Kleinschmidt & Jaeger, 2012). Later work showed that this reversal is predicted by distributional learning models (Kleinschmidt & Jaeger, 2015).

to 222 trials of exposure input. As is the norm for DL experiments, Kleinschmidt and Jaeger did not include a pre-test or incremental intermittent testing. Instead, the effect of exposure was evaluated by estimating listeners' categorization functions over the last third of the 222 trials (another common approach is to average over *all* trials, e.g., Clayards et al., 2008; Nixon et al., 2016). This revealed that listeners' categorization functions differed between exposure conditions in ways consistent with distributional learning models: the more the exposure distribution was shifted rightwards relative to each other, the more listeners' categorization functions also were shifted in the same direction.

The present work extends these findings in three ways. First, we demonstrate gradient *incremental* adaptation towards the exposure distribution. We found that the direction of the shift of the /d/ and /t/ category means in the exposure input correctly predicted the relative ordering of listeners' PSEs in Test 1-4. We also found that shifts in category means of larger magnitude (+40 vs. baseline compared to +10 vs. baseline) yielded larger shifts in listeners' PSE. Second, these changes in listeners' categorization functions were observed for natural-sounding speech that followed the types of heterogeneous distributions of phonetic cues listeners would be likely to experience during everyday speech perception. The use of natural-sounding stimuli contrasts with the type of clearly robotic sounding speech used in most previous DL experiments (for notable exceptions, see Chládková et al., 2017; Theodore & Monto, 2019). Robotic speech which might lead listeners to adopt different strategies than they would normally use. In particular, such speech provides a clear signal to listeners that some of their expectations about typical speech inputs are unlikely to extend to the current input, which might inflate listeners' readiness to adapt their perception. Similarly, our use of naturalistic *phonetic distributions* contrasts with the use of neatly designed exposure distributions that are perfectly symmetric around their mean, or that in other ways differ from the types of phonetic distributions listeners would experience in real life (for a notable exception, Chládková et al., 2017).

Third and finally, the inclusion of a pre-exposure test allowed us to see how prior expectations (prediction 1) and exposure inputs (prediction 2a,b) *jointly* explained the direction and magnitude of changes in listeners' categorization functions. We found that the direction of changes in listeners' PSEs *relative to pre-exposure* was predicted by the direction of the shift in

/d/ and /t/ distributions relative to their distributions in prior experience (Figure 6). This joint effect of prior expectations and exposure input is predicted by distributional learning accounts that explain adaptive speech perception as incremental integration of listeners' prior expectations—based on previously experienced speech input—and the statistics of the exposure input. This includes rational theories of adaptive speech perception (e.g., the ideal adaptor framework, Kleinschmidt & Jaeger, 2015, 2016), some theories of normalization (e.g., the probabilistic sliding template model, Nearey & Assmann, 2007), as well as episodic (Goldinger, 1998) and exemplar models (Johnson, 1997). Other distributional learning accounts are in principle compatible with our finding but would need to be expanded to incorporate prior expectations and the processes that integrate those expectations with new exposure input (e.g., Bejjanki et al., 2011; McMurray & Jongman, 2011; see discussion in Persson, Barreda, & Jaeger, 2024; Xie et al., 2023).

## 4.2 Prediction 4: Rapid adaptation with *diminishing returns*

Having established that exposure led to gradient changes in participants' categorization behavior, we turn to the rate with which these changes unfolded with additional exposure. The rate of change in listeners behavior is of theoretical interest for two reasons. First, it speaks to the plausibility of the same mechanisms that drive adaptive behavior in the present DL paradigm also underlie adaptive behavior during everyday speech perception (which has been found to be *very* fast, as we discuss below). Second, it speaks to the nature of the learning mechanisms that underlie adaptive speech perception. Second,

### 4.2.1 How quickly can listeners adapt their speech perception?

We found significant shifts in listeners' categorization function even after the briefest exposure tested. Exposure to 24 tokens each of shifted /d/ and /t/ was sufficient to significantly change how listeners interpreted subsequent inputs. Of note, only half of these exposure tokens labeled the intended category, the other half did not. Even when trials were labeled, labeling was indirect rather than through explicit feedback: on labeled trials, the two response options listeners saw both had the same onset stop (e.g., "din" and "dill"). Previous DL studies have assessed exposure

effects after *much* longer exposures, typically hundreds of trials (e.g., 192 trials in Harmon et al., 2019; 200 in Idemaru & Holt, 2011; 222 in Clayards et al., 2008; Kleinschmidt & Jaeger, 2016; 236 in Theodore & Monto, 2019; 456 in Nixon et al., 2016). The present results demonstrate that a fraction of the amount of exposure employed in previous studies is sufficient to elicit changes in listeners' categorization behavior.

This finding informs ongoing discussions about the type of adaptive changes observed in AA and LGPL/VGPL paradigms could plausibly arise from the same mechanisms as those observed in DL paradigms like in the present study (Baese-Berk, 2018; Bradlow & Bent, 2008; Xie et al., 2023; Zheng & Samuel, 2020). In the introduction, we mentioned findings of improved speed and accuracy of cross-modal priming after exposure to only 18 sentences from an L2-accented talker—the shortest tested exposure we are aware of (Clarke & Garrett, 2004; Xie, Weatherholtz, et al., 2018). Other AA work has more directly demonstrated that exposure changes listeners' categorization function. For example, Xie et al. (2017) found changes in listeners categorization behavior after exposure to only 30 critical L2-accented words. The directionality of these changes was consistent with distributional learning accounts of adaptive speech perception. Together with evidence from additional experiments, Xie and colleagues concluded that "listeners dynamically update their own cue-weighting functions during rapid phonetic adaptation to foreign accents, and critically over much shorter time span[s] than shown in previous studies of second language phoneme learning" (p. 215). By demonstrating that DL paradigms can elicit qualitatively similar changes with similarly little exposure, the present study lends further plausibility to the hypothesis that these changes are driven by the same underlying mechanisms.

Some experiments on LGPL/VGPL have demonstrated effects after even less exposure, with detectable changes in listeners' categorization responses after as few as 2-4 exposures to visually or lexically labeled phonetically tokens (Cummings & Theodore, 2023; Kleinschmidt & Jaeger, 2011, 2012; Liu & Jaeger, 2018, 2019; Vroomen et al., 2007). In comparing findings across paradigms, future work should keep in mind that DL and LGPL/VGPL paradigms differ in the amount of information conveyed by each exposure token. LGPL/VGPL paradigms typically employ exposure stimuli that are a) labeled and b) auditorily maximally ambiguous—falling between the two categories that the experiment focuses on. Distributional learning accounts

predict that such stimuli should be highly informative, leading to comparatively large changes in categorization behavior. This is in line with recent findings: when stimuli in LGPL/VGPL experiments are shifted less than to the point of maximal auditory ambiguity, listeners exhibit smaller shifts in categorization behavior (Babel et al., 2019; Kleinschmidt & Jaeger, 2012; Tzeng et al., 2021; see also Cummings & Theodore, 2023).

In contrast, DL paradigms a) typically employ only unlabeled stimuli (Clayards et al., 2008) or a mixture of unlabeled and labeled stimuli (e.g., Kleinschmidt & Jaeger, 2016 and the present paradigm), and b) reflect a *distribution* of phonetic properties—ranging from more to less expected under listeners' prior expectations. This makes the speech inputs in DL paradigms more similar to what one would expect during exposure to natural accents and other cross-talker differences. But it also means that exposure tokens in DL experiments are, on average, considerably less informative than in an LGPL/VGPL experiment. Future work that aims to compare the speed of adaptive speech perception across these two paradigms should thus do so *relative to the amount of information conveyed by each exposure.*

#### 4.2.2 First fast, then slow: *diminishing returns* of exposure

Our comparisons across test blocks within each exposure condition found suggestive—but not decisive—evidence that the speed of incremental changes in listeners' PSE decreased with increasing exposure: the same amount and distribution of phonetic evidence yielded smaller *additional* changes in listeners' PSE, the more exposure blocks listeners had already experienced. To the best of our knowledge, this is the first study to report this pattern of gradiently diminishing returns.[11] A similar pattern is, however, indirectly evident in at least one other recent study. Kleinschmidt (2020, Experiment 3) re-analyzes several DL experiments originally presented in Kleinschmidt and Jaeger (2016). Since these experiments lacked incremental testing, Kleinschmidt entertained several approaches to estimating incremental exposure effects, while controlling for differences in the phonetic properties of the stimuli over which these effects were

[11] Diminishing returns from input with stationary statistics—as observed here—is not to be confused with findings that listeners might become less sensitive to *changes* in a talker's speech statistics after prolonged exposure (e.g., Kraljic & Samuel, 2011; Saltzman & Myers, 2021; but see Theodore & Monto, 2019)—a finding predicted if adaptive speech perception is an active process (Magnuson & Nusbaum, 2007) that requires change detection (Qian, Jaeger, & Aslin, 2012).

assessed. The resultis unsurprisingly much noisier than in the present study, but the pattern is compatible with the hypothesis that adaptation initially proceeded quickly, and then increasingly more slowly. Paralleling the present experiment, this is particularly evident for the more extreme exposure conditions.

Such *diminishing returns* of exposure are explicitly predicted only by some distributional learning models. This includes error-driven learning models (e.g., Davis & Sohoglu, 2020; Harmon et al., 2019; Olejarczuk et al., 2018; Sohoglu & Davis, 2016) and models of ideal information integration (ideal adaptors, Kleinschmidt & Jaeger, 2015, 2016). Prior to adaptation, an idealized listener (horizontal gray ribbons in Figure 5C) would, on average, experience larger prediction errors (mean surprisal per exposure input in baseline condition $\mathrm{E}[-\log_2 p(category|VOT, f0, vowel\ duration)] = 0.62$ bits; +10 condition = 0.21 bits; +40 condition = 0.75 bits). As listeners converge towards the distribution of /d/ and /t/ in the exposure condition, they should experience increasingly smaller prediction errors (or equivalently: less new information) processing the exposure tokens. An idealized learner that has fully converged against the exposure distributions (colored lines in Figure 5C) would, on average, experience only 0.06 bits of surprisal per exposure input.

Models of adaptive speech perception that predict adaptation to be a positive monotonic function of the prediction error, thus offer an explanation for the diminishing returns of exposure observed in the present study. Of note, they do so without introducing arbitrary changes in learning rates or other parameters: it is the decrease in additional information gained from additional exposure that drives the decreasing rate of change in listeners' behavior. If the pattern of diminishing returns is replicated in future work, this would raise questions as to whether similar predictions follow from other distributional learning accounts (e.g., C-CuRE normalization, McMurray & Jongman, 2011; exemplar models, Johnson, 1997; DNNs, Magnuson et al., 2020). If, on the other hand, future tests reliably fail to replicate these findings, this would constitute a serious challenge to models that predict adaptation to be sensitive to the prediction error.

### 4.3 Prediction 3: Constraints on the early moments of adaptive speech perception?

At first glance, it would be tempting to interpret the 'diminishing returns' discussed in the previous section as evidence that listeners have converged against the exposure distributions in the input—i.e., that adaptation has successfully completed. In particular, we found that there was at best anecdotal evidence that the final exposure block had *any* additional effect (cf. hypothesis tests in Table 3). However, the comparison against idealized learners revealed that listeners had *not* actually successfully learned the phonetic distributions of the unfamiliar talker. Rather, listeners appear to have converged prematurely, long before they achieved the theoretically possible recognition accuracy.

If confirmed by additional data, premature convergence would constitute potentially severe constraints on the adaptivity of speech perception—at least during the initial moments of exposure before additional memory and learning mechanisms are engaged during, e.g., sleep (Earle & Myers, 2014; for discussion, see Fenn, Nusbaum, & Margoliash, 2003; Xie, Earle, & Myers, 2018). This would introduce a novel data point that theories of speech perception need to account for: while existing theoretical frameworks can accommodate both constraints, either constraint would substantially narrow the space of plausible models. Existing *models* of adaptive speech perception, for example, do not seem to predict these effects. Given the potential theoretical relevance of these findings, we present an additional post-hoc test to further characterize changes in participants' categorization functions. We then discuss potential explanations.

We fit a distributional learning model to listeners' data. This allows us to compare changes in listeners' behavior to those expected under an ideal—as opposed to ideal*ized*—learner.[12] Specifically, we used an ideal adaptor model (Kleinschmidt & Jaeger, 2015, 2016). The ideal adaptor defines a model of ideal information integration, updating prior beliefs about phonetic representations based on the exposure data. Univariate instances of this model have previously achieved high accuracy in modeling changes in listeners categorization functions after

[12] Following Qian, Jaeger, and Aslin (2016), we use the term *idealized* learner for models that are given privileged access to information that listeners have to infer (e.g., the correct exposure statistics of the experiment), and the term *ideal* learner for models that implement a theory of ideal information integration (for a list of assumptions made by the ideal adaptor, see appendix of Kleinschmidt & Jaeger, 2015).

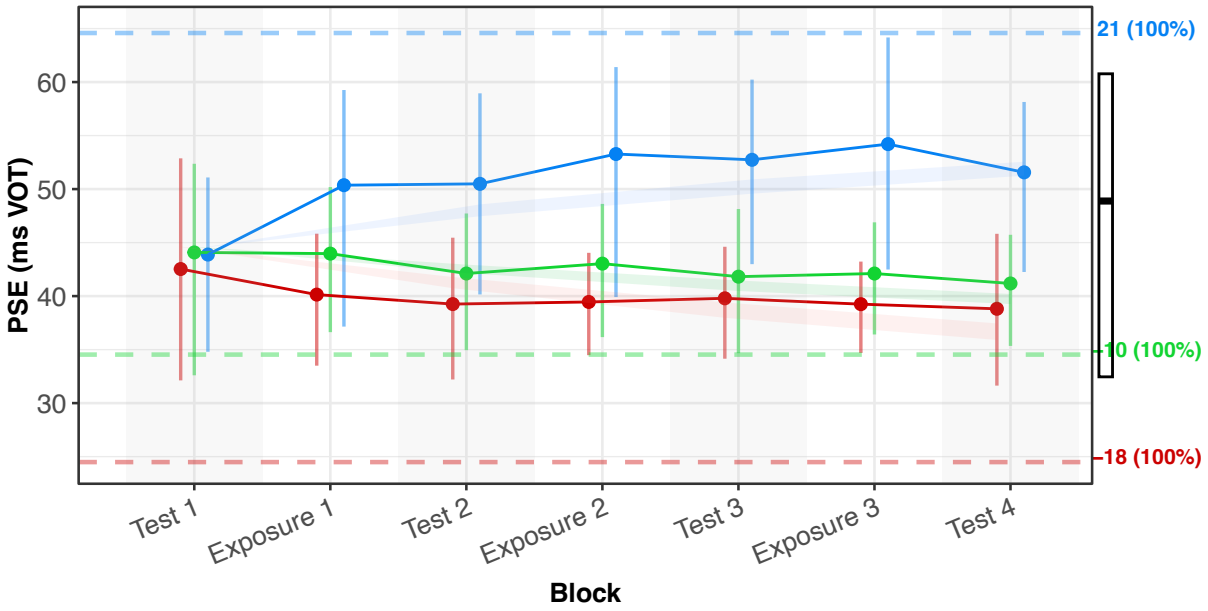


*Figure 8.* Comparison of shifts in listener PSEs over the first four test blocks against that expected by an ideal adaptor (a distributional learning model that describes ideal information integration, Kleinschmidt & Jaeger, 2015). Point ranges and dashed horizontal lines are identical to those in Figure 5C. Colored ribbons indicate the 95%-CI for the PSEs predicted by the ideal adaptor fit to listeners' responses. The vertical bar to the right of the panel indicates the range of talker-specific PSEs an idealized listener might have experienced in previous exposure: the mean and 2.5%-to-97.5% quantile range of talker-specific PSEs derived from Bayesian ideal observers fit to all talkers in the phonetic database of isolated word productions presented in Figure 6. We return to this range when discussing possible explanations for mismatches between the ideal adaptor and participants' responses.

LGPL/VGPL (for /s/-/ʃ/, Cummings & Theodore, 2023; /b/-/d/, Kleinschmidt & Jaeger, 2011, 2012) or DL exposure (for /b/-/p/, Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016; /g/-/k/, Theodore & Monto, 2019). We extended the model to multivariate categories over VOT, f0, and vowel duration, and fit it to listeners' responses in our experiment (for details, see SI §7).

Unlike the psychometric model we used for analysis, a distributional learning model specifies how information is integrated across exposure trials. As a consequence, listeners' responses during different test blocks and across different exposure conditions jointly constrain the parameters of model (due to assumption of ideal information integration within and across exposure blocks). This substantially decreases the degrees of freedom the model has to fit listeners' responses. For example, the ideal adaptor we fit to listeners' responses in the first four

test blocks has a total of 3 parameters (lapse rate, and confidence in the prior means and covariance matrices of /d/ and /t/; for details, see SI §7), compared to 24 population-level ('fixed') and 336 group-level ('random') effect parameters for the psychometric mixed-effect model fit to the same data. For the present data, the ideal adaptor model nevertheless predicts the proportion of listeners' "t"-responses with a high $R^2 = 96\%$, comparable to that of the psychometric model ($R^2 = 97\%$). This suggests that the ideal adaptor provides a decent explanation for participants' behavior with substantially fewer degrees of freedom. We leave further analysis of this model to future work. Here we focus on what the comparison of listeners behavior against the ideal adaptor can tell us about the incremental build-up of adaptation.

Despite the good quantitative fit, the ideal adaptor exhibits a systematic deficiency in explaining participants' responses. Figure 8 shows the PSEs inferred from the ideal adaptor, along with listeners' PSEs (the same data as in Figure 5C but focused on blocks up to Test 4). Notably, the ideal adaptor *under*-predicts changes in listeners' PSEs in early blocks—a tendency that is most evident for the exposure conditions with more extreme shifts (baseline and +40). Similarly, though less obvious, the ideal adaptor *over*-predicts listeners' PSE in later blocks (see the baseline and +40 conditions in Test 4).

To understand this pattern, it is helpful to recognize that rational models of distributional learning describe listeners' expectations after exposure as the weighted mean of prior expectations and the observed exposure statistics (Kleinschmidt & Jaeger, 2015). This means that the speed with which a learner adapts at the start of exposure is expected to be predictive of how fast a learner adapts after additional exposure. Put differently, ideal adaptors constrain the rate at which PSEs can change across exposure trials. This means that an ideal adaptor can *either* explain the fact that listeners' PSEs changed rapidly at the start of exposure *or* that there seem to be little to no changes in listeners' PSE after the second exposure block. But it cannot explain both aspects of our data. It thus appears that participants indeed converged prematurely in their behavior, relative to what would be expected under the specific model of ideal information integration that we employed.

Explanations for this result can be grouped into two classes. One type of explanation takes at face value that the 'flattening off' of the changes in participants' behavior reflects a form of

premature convergence, and that it reflects a previously unrecognized constraint on adaptive speech perception. Alternatively, the observed patterns might be due to methodological artifacts, or a failure to take into account aspects of our design in reasoning through theories of adaptive speech perception. Under this interpretation, the present data do not constitute evidence against the hypothesis that adaptive speech perception can achieve full learning. We discuss these two classes of explanations in turn, beginning with explanations that have consequences for theories of adaptive speech perception.

One explanation for premature convergence is that adaptive speech perception—or at least its earliest moments—is *not* the result of distributional learning. This would require the formulation of, as of yet unspecified, alternative mechanisms. It would also require explanations for the properties of adaptive speech perception that *do* receive explanations under the hypothesis that adaptive speech perception involves distributional learning, including the properties we have demonstrated above.

There are, however, alternative explanations that do not throw out the proverbial baby with the bathwater. Under the first of these explanations, adaptive speech perception *is* driven by distributional learning, just not *unconstrained* distributional learning. This is the hypothesis we consider most plausible at this point (while maintaining considerable uncertainty about it), and so we elaborate on it in some detail. Specifically, it is possible that rapid adaptation during the earliest moments of encountering an unfamiliar talker is limited to the selection of (mixtures of) previously experienced talkers.

To appreciate this idea, we introduce a distinction proposed by Xie and colleagues (2018, pp. 2028–2029): the difference between *model learning* and *model selection*. Model learning refers to the idea that listeners learn new phonetic category representations for the unfamiliar talker. This is the idea implemented in the ideal adaptor *model*—not to be confused with the more general *theory*—that we used above to derive the prediction in Figure 8. Similarly, the idea that listeners store speech episodes (Goldinger, 1998, 2007) or exemplars (Hay, 2018; Johnson, 1997) from the unfamiliar talker, and then use these to categorize subsequent speech from that talker is a form of model learning. Critically, as such learning continues, listeners should increasingly come to reflect categorization behavior that is based on the phonetic distributions of the new

talker—contrary to what we seem to observe.

Model selection, on the other hand, refers to the idea that listeners select between different *previously experienced* models. In the case of speech perception, the models being selected between are talker- or talker group-specific phonetic representations (e.g., idiolects, dialects, sociolects, etc., as reviewed in Pajak et al., 2016). Each of these models specifies a mapping from phonetic cues to categories. Model selection, too, can be seen as a form of distributional learning, as the incremental reweighting of different models based on both top-down (contextual) and bottom-up (acoustic) cues to talker identity (Kleinschmidt et al., 2015, pp. 180–182). This reweighting allows listeners to adapt to unfamiliar input by upweighting previously learned models that more accurately categorize speech from the new talker. Unlike model learning, however, the flexibility afforded by model selection is limited, and strongly constrained by previous experience. Specifically, model selection alone would only allow listeners to adapt their behavior *up to the most extreme* previously stored model. This is why Kleinschmidt and Jaeger hypothesized that human speech adaptation might draw on both model selection and model learning in order to strike a trade-off between stability and flexibility (Kleinschmidt et al., 2015, Equations 24-25, p. 181). But what if listener only or primarily relied on model selection? Xie, Weatherholtz, et al. (2018) proposed that there are reasons to believe that at least the earliest moments of adaptive speech perception might be primarily driven by model selection, and that model learning might depend on slower neural mechanisms, such as memory consolidation of new exemplars during sleep (for relevant discussion, see Estes, 1986; Fenn & Hambrick, 2013; Xie, Earle, et al., 2018).

To assess whether primary reliance on model selection might be a plausible explanation for premature convergence, we again used the phonetic database of /d/ and /t/ productions shown in Figure 6 (Chodroff & Wilson, 2018). This time, we fit separate ideal observers to all talkers in that database to obtain the predicted PSEs for each of those talkers. This allowed us to estimate the range of talker-specific PSEs that a typical L1 listener of US English might have experienced throughout their life prior to our experiment. This serves as an estimate of the range of PSEs that a listener would be expected to accommodate based on model selection alone. This range is indicated on the righthand-side of Figure 8. It provides a decent qualitative fit to the range of adaptive changes in the PSEs that listeners in our experiment accommodated.

Model selection—listeners selecting between previously learned phonetic representations—thus provides one plausible explanation for premature convergence. This explanation makes a prediction that can be easily tested in future work: even substantially longer exposure to, for example, a few hundred exposure trials should still not result in convergence against the behavior of an idealized learner—at least as long as exposure is limited to a single day without intermittent sleep. A separate question for future research would be whether listeners can overcome the initial premature convergence with repeated exposure over multiple days, as hypothesized in Xie, Weatherholtz, et al. (2018).

There are, of course, other potential explanations for premature convergence. One type of explanation appeals to methodological confounds. For example, we considered whether premature convergence could be a trivial result of priors we used in fitting the psychometric mixed-effects model. Following standards in the literature, we employed a weakly regularizing priors Student $t$ prior for all population-level effects. This prior favors small coefficient estimates, regularizing the estimated shifts in PSEs towards zero. As this regularization is particularly strong for more extreme shifts in the PSE, it is theoretically possible that our priors caused the psychometric model to 'hallucinate' premature convergence. This would make these findings artifacts of our data analysis approach, rather than findings of theoretical interest. Given how weakly regularizing the priors we employed were, this explanation struck us as unlikely: even the largest estimated shifts were well within the 95% highest density interval of the prior. Still, we decided to address this possibility more directly. We refit the psychometric model once with a substantially weaker prior (SD of Student $t = 5$) and once with an even weaker uniform prior. In both cases, results changed only numerically and premature convergence was still observed.

Other possible explanations appeal to assumptions we made for the idealized learner who has fully learned the exposure distributions, and how these assumptions mismatch the information that is available to listeners. Under this explanation, what appears as premature convergence in reality reflects adequate convergence against what would be expected from an idealized learner that has access to the same information as listeners. For example, our idealized learner models have perfectly learned the statistics of the exposure stimuli, while ignoring all test stimuli. Listeners, however, might learn even from unlabeled test stimuli. Indeed, the effects of

repeated testing without intermittent exposure (Test 4-6) suggests as much.

Critically, test tokens had by design identical phonetic properties across all exposure conditions. Inevitably then, the location of test tokens relative to the exposure tokens *differed* between conditions. If listeners integrate test tokens into their estimate of the talker's accent, this might explain premature convergence: at the end of Test 4, 36 out of 180 trials (20%) that participants had experienced were test tokens.[13] While it is difficult to evaluate this explanation without a specific model of how listener learn from unlabeled tokens, one consideration suggests that it is not sufficient to explain our data. To estimate how much learning test tokens alone would support in the different conditions, we calculated the surprisal of the test token under the idealized learners. For an idealized learner that has *fully* learned the exposure distribution (cf. colored dashed lines in Figure 5C), test stimuli would convey about the same amount of surprisal in the baseline and +10 conditions (both $\mathrm{E}[-\log_2 p(VOT|\text{idealized learner})] = 5.7$ bits), compared to larger surprisal in the + 40 condition (7.2 bits). At least based on these general considerations, learning from test tokens alone would thus predict even earlier premature convergence in the +40 conditions, compared to the other two conditions—the opposite of what we observed. Future work can further address this question by developing and applying unsupervised adaptation models to our data (e.g., Harmon et al., 2019; Olejarczuk et al., 2018; Yan & Jaeger, 2018). Future work could more decisively address this alternative explanation by replicating our experiment while using test tokens that are placed identically *relative to the exposure distributions*.

## 4.4 Limitations and future directions

The present study set out to investigate incremental adaptation to a single talker, whose pronunciations were shifted relative to listeners' prior expectations. The conclusions we have discussed so far should be interpreted in light of several limitations of the approach we took. First, our experiment investigated incremental adaptation to a single female talker's productions of syllable-initial /d/-/t/ by L1-US English listeners. In particular, our exposure conditions

[13] If listeners adapt over a moving time-window (rather than over all inputs from a talker), or in other ways weight more recent information more strongly, this would further increase the relative impact of test tokens on listeners' categorization responses during test.

shifted the distribution of VOT and, by extension, f0 and vowel duration. As adaptive speech perception—including its generalization across categories, phonetic contexts, and talkers—can depend on the exposure talker, the phonetic contrast or the phonetic cues it involves (e.g., Eisner & McQueen, 2005; Kraljic & Samuel, 2007; Mitterer et al., 2013; Xie, Liu, et al., 2021; Xie et al., 2017), future work is necessary to assess how general the present findings are.

Second, the present study shares with other DL paradigms that a small number of minimal pair items was repeated many times, with only minimal phonetic differences embedded in otherwise constant phonetic contexts (e.g., the vowel following /d/-/t/ was always the same), and presented in isolation. This sacrifice of ecological validity was motivated by our goal to test stronger predictions about the direction and relative magnitude of effects (rather than merely the existence of effects). It does, however, mean that the speech input that participants experienced in the experiment differed from everyday encounters with unfamiliar talkers: listeners typically experience *connected* speech from unfamiliar talkers, which tends to be produced with faster speech rates and comes with additional segmentation challenges; while the same phonetic contrast might appear many times, it will not necessarily appear in the same phonetic context, least of all in the same word; and the speech of talkers with unfamiliar accents often deviate from listeners' expectations in more than a single segmental contrast. Comparatively little is known about adaptive speech perception under such more common conditions (even AA studies have typically focused on short isolated sentences, Bradlow & Bent, 2008; **clark-garrett2024?**).

Third, as already mentioned, some of the tests were conducted post-hoc, and thus should be interpreted with caution. In particular, future experiments with longer exposure would provide more decisive tests of the explanations we offered for premature convergence. Additional data from future applications of the incremental exposure-test paradigm with different phonetic shifts will also facilitate stronger quantitative tests of models of adaptive speech perception (in the spirit of Guest & Martin, 2021; Xie et al., 2023; Yarkoni & Westfall, 2017). In a recent review of the field, Xie et al. (2023) demonstrated that the signature findings of some of the most popular paradigms in adaptive speech perception do not distinguish between radically different theoretical accounts. Qualitative improvements in speech recognition can be explained by mechanisms ranging from early pre-linguistic perceptual normalization, changes in the representations of

phonetic categories, or upstream changes in decision-making. Xie and colleagues concluded that the effective comparisons of these theories will require quantitative data sets that constrain the way in which listeners' categorization behavior changes depending on the amount and nature of the input. The use of incremental testing and multiple exposure conditions with different phonetic shifts—as explored in the present study—provides such data.

# 5 Conclusions

Research on adaptive changes in speech perception has made great strides since foundational work in the 90s and 00s. Now that the existence of adaptive changes in speech perception is no longer in question, recent reviews of the field have emphasized the need to develop novel paradigms that can inform the functional relation between exposure inputs and changes in listeners' perception (Schertz & Clare, 2020; Xie et al., 2023). The present study is a response to this call. We set out to more clearly characterize the incremental unfolding of adaptation to changes in the realization of a simple two-way phonetic contrast. This allowed us to assess previously untested predictions of distributional learning accounts of adaptive speech perception.

In line with these theories, we found that listeners initially draw on prior experience with other talkers to recognize input from the unfamiliar talker. With increasing exposure, listeners then adapted their categorization responses, improving recognition accuracy. The incremental unfolding of these changes followed the prediction of distributional learning accounts—in particular, accounts that predict changes in listeners' perception to depend on the prediction error (or the amount of new information) associated with each new exposure input. Finally, we found suggestive evidence that adaptivity, at least during the earliest moments of exposure, is constrained: while adaptation was rapid, it also slowed down and seemed to converge against a stable state long before listeners approached the recognition accuracy expected from an idealized learner. The specific nature of these constraints strikes us as an important target for future work, as they potentially impose novel constraints on theories of adaptive speech perception.

Our findings were facilitated by both Bayesian psychometric mixed-effects analysis (Kuss et al., 2005; e.g., Prins, 2019a) and normative models of adaptive speech perception (Kleinschmidt &

Jaeger, 2015). We extended the former to fit a single model across all participants, while correcting for participant-specific lapse rates and while modeling block-by-block changes in participants' psychometric functions. While such models used to require expensive software, freely available software has substantially lowered the entry cost for such approaches (e.g., `R`, R Core Team, 2023;`brms` Bürkner, 2017). Similarly, there are now R libraries that facilitate the fitting and evaluation of ideal observers and adaptors (T. Florian Jaeger, n.d.; **beliefupdatr?**). The R markdown available on OSF provides a starting point for other researchers interested in either type of model.

# 6 References


Adams, T. L., Li, Y., & Liu, H. (2020). A replication of beyond the turk: Alternative platforms for crowdsourcing behavioral research–sometimes preferable to student groups. *AIS Transactions on Replication Research*, *6*(1), 15.

Albert, D. A., & Smilek, D. (2023). Comparing attentional disengagement between prolific and MTurk samples. *Scientific Reports*, *13*(1), 20574.

Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *The Journal of the Acoustical Society of America*, *106*(4), 2031–2039.

Apfelbaum, K. S., & McMurray, B. (2015). Relative cue encoding in the context of sophisticated models of categorization: Separating information from categorization. *Psychonomic Bulletin & Review*, *22*, 916–943.

Aust, F., & Barth, M. (2023). *papaja: Prepare reproducible APA journal articles with R Markdown*. Retrieved from https://github.com/crsh/papaja

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

Bache, S. M., & Wickham, H. (2022). *Magrittr: A forward-pipe operator for r*. Retrieved from https://CRAN.R-project.org/package=magrittr

Baese-Berk, M. (2018). Perceptual learning for native and non-native speech. In *Psychology of learning and motivation* (Vol. 68, pp. 1–29). Elsevier.

Barth, M. (2023). *tinylabels: Lightweight variable labels*. Retrieved from https://cran.r-project.org/package=tinylabels

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. *PloS One*, *6*(5), e19812.

Bent, T., & Baese-Berk, M. (2021). Perceptual learning of accented speech. *The Handbook of Speech Perception*, 428–464.

Bicknell, K., Bushong, W., Tanenhaus, M. K., & Jaeger, T. F. (under review). *Maintenance of subcategorical information during speech perception: Revisiting misunderstood limitations.*

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer. Version 6.2. 12*.

Bolker, B., & Robinson, D. (2022). *Broom.mixed: Tidying methods for mixed models*. Retrieved from https://CRAN.R-project.org/package=broom.mixed

Bradlow, A. R., Bassard, A. M., & Paller, K. A. (2023). Generalized perceptual adaptation to second-language speech: Variability, similarity, and intelligibility. *The Journal of the Acoustical Society of America*, *154*(3), 1601–1613.

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

Burchill, Z. (2023). *The reliability of standard reading time analyses and understanding the nature of maintained information in speech processing* (PhD thesis). University of Rochester.

Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Maintaining information about speech input during accent adaptation. *PloS One*, *13*(8), e0199358.

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

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5), 1–54. https://doi.org/10.18637/jss.v100.i05

Bushong, W., & Jaeger, T. F. (2019). Dynamic re-weighting of acoustic and contextual

cues in spoken word recognition. *The Journal of the Acoustical Society of America*, *146*(2), EL135–EL140.

Bushong, W., & Jaeger, T. F. (under review). *Maintenance of subcategorical representations in spoken word recognition is modulated by recent experience.*

Campitelli, E. (2024). *Ggnewscale: Multiple fill and colour scales in 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggnewscale

Chang, W. (2023). *Webshot: Take screenshots of web pages*. Retrieved from https://CRAN.R-project.org/package=webshot

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.

Chodroff, E., & Wilson, C. (2014). Burst spectrum as a cue for the stop voicing contrast in american english. *The Journal of the Acoustical Society of America*, *136*(5), 2762–2772.

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).

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented english. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658.

Clayards, M. (2017). Individual talker and token covariation in the production of multiple cues to stop voicing. *Phonetica*, *75*(1), 1–23.

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

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

Cummings, S. N., & Theodore, R. M. (2023). Hearing is believing: Lexically guided

perceptual learning is graded to reflect the quantity of evidence in speech input. *Cognition*, *235*, 105404.

Davis, M. H., & Sohoglu, E. (2020). Three functions of prediction error for bayesian inference in speech perception. *The Cognitive Neurosciences*, 177–189.

Dmitrieva, O., Llanos, F., Shultz, A. A., & Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in spanish and english. *Journal of Phonetics*, *49*, 77–95.

Docherty, G. J. (1992). *The timing of voicing in british english obstruents*. Walter de Gruyter.

Earle, F. S., & Myers, E. B. (2014). Building phonetic categories: An argument for the role of sleep. *Frontiers in Psychology*, *5*, 108124.

Eddelbuettel, D., & Balamuta, J. J. (2018). Extending R with C++: A Brief Introduction to Rcpp. *The American Statistician*, *72*(1), 28–36. https://doi.org/10.1080/00031305.2017.1375990

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. https://doi.org/10.18637/jss.v040.i08

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

Eisner, F., Melinger, A., & Weber, A. (2013). Constraints on the transfer of perceptual learning in accented speech. *Frontiers in Psychology*, *4*, 148.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, *18*(4), 500–549.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752.

Fenn, K. M., & Hambrick, D. Z. (2013). What drives sleep-dependent memory consolidation: Greater gain or less loss? *Psychonomic Bulletin & Review*, *20*, 501–506.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature*, *425*(6958), 614–616.

Francis, A. L., Ciocca, V., & Ching Yu, J. M. (2003). Accuracy and variability of acoustic

measures of voicing onset. *The Journal of the Acoustical Society of America*, *113*(2), 1025–1032.

Frick, H., Chow, F., Kuhn, M., Mahoney, M., Silge, J., & Wickham, H. (2023). *Rsample: General resampling infrastructure*. Retrieved from https://CRAN.R-project.org/package=rsample

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*(15), 2865–2873.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). *A weakly informative default prior distribution for logistic and other regression models*.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), 555.

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.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251.

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. *Proceedings of the 16th International Congress of Phonetic Sciences*, 49–54. Citeseer.

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*(3), 1–25. Retrieved from https://www.jstatsoft.org/v40/i03/

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

Harmon, Z., Idemaru, K., & Kapatsinski, V. (2019). Learning mechanisms in cue reweighting. *Cognition*, *189*, 76–88.

Hay, J. (2018). Sociophonetics: The role of words, the role of context, and the role of words in context. *Topics in Cognitive Science*, *10*(4), 696–706.

Henry, L., & Wickham, H. (2024). *Rlang: Functions for base types and core r and 'tidyverse' features*. Retrieved from https://CRAN.R-project.org/package=rlang

Henry, L., Wickham, H., & Chang, W. (2022). *Ggstance: Horizontal 'ggplot2' components*. Retrieved from https://CRAN.R-project.org/package=ggstance

Hitczenko, K., & Feldman, N. H. (2016). Modeling adaptation to a novel accent. *Proceedings of the Annual Conference of the Cognitive Science Society*.

Hörberg, T., & Jaeger, T. F. (2021). A rational model of incremental argument interpretation: The comprehension of swedish transitive clauses. *Frontiers in Psychology*, *12*, 674202.

Hugh-Jones, D. (2024). *Latexdiffr: Diff TeX, 'rmarkdown' or 'quarto' files using the 'latexdiff' utility*. Retrieved from https://CRAN.R-project.org/package=latexdiffr

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939.

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

Jaeger, T. Florian. (n.d.). *MVBeliefUpdatr: Fitting, summarizing, and visualizing multivariate gaussian ideal observers and adaptors*. Retrieved from https://github.com/hlplab/MVBeliefUpdatr

Jaeger, T. Florian. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Johnson, K. (1997). Speech perception without speaker normalization. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–146). San Diego: Academic Press.

Kay, M. (2023). *tidybayes: Tidy data and geoms for Bayesian models*. https://doi.org/10.5281/zenodo.1308151

Kirby, J. P., & Ladd, D. R. (2016). Effects of obstruent voicing on vowel F0: Evidence from "true voicing" languages. *The Journal of the Acoustical Society of America*, *140*(4), 2400–2411.

Kleinschmidt, D. F. (2020). *What constrains distributional learning in adults?*

Kleinschmidt, D. F. (2023). *Supunsup: Supervised/unsupervised adaptation analysis*.

Kleinschmidt, D. F., Burchill, Z., Bushong, W., Karboga, G., Jaeger, F., Liu, L., … Xie, X. (2021). *JSEXP*. Retrieved from https://github.com/hlplab/JSEXP

Kleinschmidt, D. F., & Jaeger, T. F. (2011). A bayesian belief updating model of phonetic recalibration and selective adaptation. *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 10–19.

Kleinschmidt, D. F., & Jaeger, T. F. (2012). A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*.

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.

Kleinschmidt, D. F., & Jaeger, T. F. (2016). What do you expect from an unfamiliar talker? *CogSci*.

Kleinschmidt, D. F., Raizada, R. D., & Jaeger, T. F. (2015). Supervised and unsupervised learning in phonetic adaptation. *CogSci*.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178.

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

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.

Kraljic, T., & Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition*, *121*(3), 459–465.

Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, *23*(6), 1681–1712.

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2018). Effects of distributional information on categorization of prosodic contours. *Psychonomic Bulletin & Review*, *25*, 1153–1160.

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, *5*(5), 8–8.

Lancia, L., & Winter, B. (2013). The interaction between competition, learning, and habituation dynamics in speech perception. *Laboratory Phonology*, *4*(1), 221–257.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001.

Liao, Y. (2019). *Linguisticsdown: Easy linguistics document writing with r markdown*. Retrieved from https://CRAN.R-project.org/package=linguisticsdown

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

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.

Maechler, M. (2023). *Diptest: Hartigan's dip test statistic for unimodality - corrected*. Retrieved from https://CRAN.R-project.org/package=diptest

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391.

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabi, M., et al.others. (2020). EARSHOT: A minimal neural network model of incremental human speech recognition. *Cognitive Science*, *44*(4), e12823.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, *21*(3), 398–421.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, *97*(2), 225.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

McClelland, J. L., Thomas, A. G., McCandliss, B. D., & Fiez, J. A. (1999). Understanding failures of learning: Hebbian learning, competition for representational space, and

some preliminary experimental data. *Progress in Brain Research*, *121*, 75–80.

McCloy, D. R. (2016). *phonR: Tools for phoneticians and phonologists*.

McMurray, B., Cole, J. S., & Munson, C. (2011). Features as an emergent product of computing perceptual cues relative to expectations. *Where Do Features Come from*, 197–236.

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.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126.

Mikuteit, S., & Reetz, H. (2007). Caught in the ACT: The timing of aspiration and voicing in east bengali. *Language and Speech*, *50*(2), 247–277.

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

Müller, K. (2020). *Here: A simpler way to find your files*. Retrieved from https://CRAN.R-project.org/package=here

Müller, K., & Wickham, H. (2023). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Nearey, T. M., & Assmann, P. F. (2007). Probabilistic "sliding template" models for indirect vowel normalization. In M. J. Solé, P. S. Beddor, & M. Ohala (Eds.), *The Journal of the Acoustical Society of America* (pp. 246–269).

Nixon, J. S., Rij, J. van, Mok, P., Baayen, R. H., & Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: Eye movement evidence from cantonese segment and tone perception. *Journal of Memory and Language*, *90*, 103–125.

Norris, D., & Cutler, A. (2021). More why, less how: What we need from models of cognition. *Cognition*, *213*, 104688.

Norris, D., & McQueen, J. M. (2008). Shortlist b: A bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357.

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

Olejarczuk, P., Kapatsinski, V., & Baayen, R. H. (2018). Distributional learning is error-driven: The role of surprise in the acquisition of phonetic categories. *Linguistics Vanguard*, *4*(s2), 20170020.

Ooms, J. (2024). *Magick: Advanced graphics and image-processing in r*. Retrieved from https://CRAN.R-project.org/package=magick

Pajak, B., Fine, A. B., Kleinschmidt, D. F., & Jaeger, T. F. (2016). Learning additional languages as hierarchical probabilistic inference: Insights from first language processing. *Language Learning*, *66*(4), 900–944.

Pajak, B., & Levy, R. (2012). Distributional learning of L2 phonological categories by listeners with different language backgrounds. *Proceedings of the 36th Boston University Conference on Language Development*, *2*, 400–413. Cascadilla Press Somerville, MA.

Pedersen, T. L. (2024a). *Ggforce: Accelerating 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggforce

Pedersen, T. L. (2024b). *Patchwork: The composer of plots*. Retrieved from https://CRAN.R-project.org/package=patchwork

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163.

Peirce, J. W. (2007). PsychoPy—psychophysics software in python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13.

Persson, A., Barreda, S., & Jaeger, T. F. (2024). *Comparing accounts of formant normalization against US english listeners' vowel perception*.

Persson, A., & Jaeger, T. F. (2023). Evaluating normalization accounts against the dense vowel space of central swedish. *Frontiers in Psychology*, *14*, 1165742.

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(2), 297.

Prins, N. (2011). The psychometric function: Why we should not, and need not, estimate

the lapse rate. *Journal of Vision*, *11*(11), 1175–1175.

Prins, N. (2019a). Hierarchical bayesian modeling of the psychometric function (and an example application in an experiment on correspondence matching in long-range motion). *Journal of Vision*, *19*(10), 287b–287b.

Prins, N. (2019b). Too much model, too little data: How a maximum-likelihood fit of a psychometric function may fail, and how to detect and avoid this. *Attention, Perception, & Psychophysics*, *81*, 1725–1739.

Qian, T., Jaeger, T. F., & Aslin, R. N. (2012). Learning to represent a multi-context environment: More than detecting changes. *Frontiers in Psychology*, *3*, 228.

Qian, T., Jaeger, T. F., & Aslin, R. N. (2016). Incremental implicit learning of bundles of statistical patterns. *Cognition*, *157*, 156–173.

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 539.

RStudio Team. (2020). *RStudio: Integrated development environment for r*. Boston, MA: RStudio, PBC. Retrieved from http://www.rstudio.com/

Saltzman, D., & Myers, E. (2021). Listeners are initially flexible in updating phonetic beliefs over time. *Psychonomic Bulletin & Review*, *28*, 1354–1364.

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

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2016). Individual differences in perceptual adaptability of foreign sound categories. *Attention, Perception, & Psychophysics*, *78*, 355–367.

Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production.

*Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(2), e1521.

Schmale, R., Cristia, A., & Seidl, A. (2012). Toddlers recognize words in an unfamiliar accent after brief exposure. *Developmental Science*, *15*(6), 732–738.

Schuster, S. (2020). *Praat: Doing phonetics by computer [computer program]*. Stanford, CA: Interactive Language Processing Lab Stanford. Retrieved from https://docs.proliferate.alps.science/en/latest/contents.html

Schütt, H., Harmeling, S., Macke, J., & Wichmann, F. (2015). Psignifit 4: Pain-free bayesian inference for psychometric functions. *Journal of Vision*, *15*(12), 474–474.

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.

Sohoglu, E., & Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing prediction error. *Proceedings of the National Academy of Sciences*, *113*(12), E1747–E1756.

Sonderegger, M., Stuart-Smith, J., Knowles, T., Macdonald, R., & Rathcke, T. (2020). Structured heterogeneity in scottish stops over the twentieth century. *Language*, *96*(1), 94–125.

Tan, M., Xie, X., & Jaeger, T. F. (2021). Using rational models to interpret the results of experiments on accent adaptation. *Frontiers in Psychology*, 4523.

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.

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

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.

Vaughan, D., & Dancho, M. (2022). *Furrr: Apply mapping functions in parallel using futures*. Retrieved from https://CRAN.R-project.org/package=furrr

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved rhat for assessing convergence of MCMC (with discussion). *Bayesian Analysis*.

Vroomen, J., & Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition*, *110*(2), 254–259.

Vroomen, J., Linden, S. van, De Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia*, *45*(3), 572–577.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Weatherholtz, K., & Jaeger, T. F. (2016). Speech perception and generalization across talkers and accents. In *Oxford research encyclopedia of linguistics*.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019). *Assertthat: Easy pre and post assertions*. Retrieved from https://CRAN.R-project.org/package=assertthat

Wickham, H. (2023a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2023b). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2023). *Purrr: Functional programming tools*. Retrieved from

https://CRAN.R-project.org/package=purrr

Wickham, H., Hester, J., & Bryan, J. (2024). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

Wickham, H., Vaughan, D., & Girlich, M. (2024). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wilke, C. O., & Wiernik, B. M. (2022). *Ggtext: Improved text rendering support for 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggtext

Winn, M. B. (2020). Manipulation of voice onset time in speech stimuli: A tutorial and flexible praat script. *The Journal of the Acoustical Society of America*, *147*(2), 852–866.

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

Xie, X., Buxó-Lugo, A., & Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition*, *211*, 104619.

Xie, X., Earle, F. S., & Myers, E. B. (2018). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience*, *33*(2), 196–210.

Xie, X., Jaeger, T. F., & Kurumada, C. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review. *Cortex*.

Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology: General*, *150*(11), e22.

Xie, X., Theodore, R. M., & Myers, E. B. (2017). More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 206.

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.

Yan, S., & Jaeger, T. F. (2018). Comparing models of unsupervised adaptation in speech perception. *The 24th Annual Conference on Architectures and Mechanisms for Language Processing (AMLaP)*. Berlin, Germany: poster.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

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.

Zheng, Y., & Samuel, A. G. (2023). Flexibility and stability of speech sounds: The time course of lexically-driven recalibration. *Journal of Phonetics*, *97*, 101222.

Zhu, H. (2024). *kableExtra: Construct complex table with 'kable' and pipe syntax*. Retrieved from https://CRAN.R-project.org/package=kableExtra

# Supplementary information

# Contents

# §1 Required software

Both the main text and these supplementary information (SI) are derived from the same R markdown document available via https://osf.io/hxcy4/. It is best viewed using Acrobat Reader. The document was compiled using `knitr` in RStudio with R:

```
##                _
## platform       aarch64-apple-darwin20
## arch           aarch64
## os             darwin20
## system         aarch64, darwin20
## status
## major          4
## minor          3.1
```

```
## year           2023
## month          06
## day            16
## svn rev        84548
## language       R
## version.string R version 4.3.1 (2023-06-16)
## nickname       Beagle Scouts
```

You will also need to download the IPA font SIL Doulos and a Latex environment like (e.g., MacTex or the R library `tinytex`).

We used the following R packages to create this document: R (Version 4.3.1; R Core Team, 2023) and the R-packages *assertthat* (Version 0.2.1; Wickham, 2019), *brms* (Version 2.20.4; Bürkner, 2017, 2018, 2021), *broom.mixed* (Version 0.2.9.4; Bolker & Robinson, 2022), *diptest* (Version 0.77.0; Maechler, 2023), *dplyr* (Version 1.1.4; Wickham, François, Henry, Müller, & Vaughan, 2023), *forcats* (Version 1.0.0; Wickham, 2023a), *furrr* (Version 0.3.1; Vaughan & Dancho, 2022), *ggforce* (Version 0.4.2; Pedersen, 2024a), *ggnewscale* (Version 0.4.10; Campitelli, 2024), *ggplot2* (Version 3.5.0; Wickham, 2016), *ggstance* (Version 0.3.6; Henry, Wickham, & Chang, 2022), *ggtext* (Version 0.1.2; Wilke & Wiernik, 2022), *here* (Version 1.0.1; Müller, 2020), *kableExtra* (Version 1.4.0; Zhu, 2024), *latexdiffr* (Version 0.2.0; Hugh-Jones, 2024), *linguisticsdown* (Version 1.2.0; Liao, 2019), *lubridate* (Version 1.9.3; Grolemund & Wickham, 2011), *magick* (Version 2.8.3; Ooms, 2024), *magrittr* (Version 2.0.3; Bache & Wickham, 2022), *papaja* (Version 0.1.1.9001; Aust & Barth, 2023), *patchwork* (Version 1.2.0; Pedersen, 2024b), *phonR* (Version 1.0.7; McCloy, 2016), *posterior* (Version 1.5.0; Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2021), *purrr* (Version 1.0.2; Wickham & Henry, 2023), *Rcpp* (Eddelbuettel & Balamuta, 2018; Version 1.0.12; Eddelbuettel & François, 2011), *readr* (Version 2.1.5; Wickham, Hester, & Bryan, 2024), *rlang* (Version 1.1.3; Henry & Wickham, 2024), *rsample* (Version 1.2.0; Frick et al., 2023), *stringr* (Version 1.5.1; Wickham, 2023b), *supunsup* (Version 0.2.0; Kleinschmidt, 2023), *tibble* (Version 3.2.1; Müller & Wickham, 2023), *tidybayes* (Version 3.0.6; Kay, 2023), *tidyr* (Version 1.3.1; Wickham, Vaughan, & Girlich, 2024), *tidyverse* (Version 2.0.0; Wickham et al., 2019), *tinylabels* (Version 0.2.4; Barth, 2023), and *webshot* (Version 0.5.5; Chang, 2023). If opened in

RStudio, the top of the R markdown document should alert you to any libraries you will need to download, if you have not already installed them. The full session information is provided at the end of this document.

### §1.1 Interested in using R markdown do create APA formatted documents that integrate your code with your writing?

A project template, including R markdown files that result in APA-formatted PDFs, is available at https://github.com/hlplab/template-R-project. Feedback welcome. We aim to help others avoid the mistakes and detours we made when first deciding to embrace literal coding to increase transparency in our projects.

## §2 Additional information about Materials

All stimuli are available as part of the OSF repository for this article.

### §2.1 Recordings

An L1-US English female talker originally from New Hampshire was recruited for recording of the stimuli. She was recorded at the Human Language Processing lab at the Brain & Cognitive Sciences Department, University of Rochester with the help of research assistant (also an L1-US English speaker). She was 23 years old at the time of recording and was judged by the research assistant to have a generic US American accent known as "general American".

Four /d-t/ minimal pairs (*dill-till*, *din-tin*, *dim-tim*, *dip-tip*) were recorded together with 20 filler words. These fillers were made up of 10 minimal or near minimal pairs with different sounds at onset. The word pairs were separated into two lists so that they would appear in separate blocks during recording. Each critical pair was repeated 8 times while the filler pairs were repeated 5 times. Word presentation was delivered with PsychoPy (Peirce, 2007) and the presentation was controlled by the researcher from a computer located outside the recording room. The order of each block was randomized such that target words never appeared consecutively. The talker was instructed to speak clearly and confidently, and to maintain a

consistent distance from the microphone.

## §2.2 Resynthesis

Separate procedures were used to generate stimuli with positive and negative VOTs. We first describe the annotation of the original recordings required by both procedures. Then we describe the two procedures, followed by the annotation approach used for the resynthesized stimuli.

### §2.2.1 Annotation of original recordings

All critical pairs of the talker's recordings were annotated for the four cues that are known to affect the perception of word-initial stop-voicing in L1-US English. Durational measurements of pre-voicing, VOT, and vowel duration were taken in addition to the mean F0 of the first 25% of the vowel duration. Annotations were done on Praat and based on both listening and inspection of the waveform and spectrogram. The annotation boundaries were made following approaches reported in prior studies (Clayards, 2017; Dmitrieva, Llanos, Shultz, & Francis, 2015; e.g. Francis, Ciocca, & Ching Yu, 2003; Kirby & Ladd, 2016) and through personal communication with trained phoneticians.

- pre-voicing (voicing during closure)
  - **start:** the first sign of periodicity in the waveform before closure release.
  - **End:** the point of closure release
- VOT
  - **start:** the point of closure release.
  - **End:** the beginning of clearly defined periodicity in the waveform and at the appearance of low frequency energy in the spectrogram.
- Vowel
  - **start:** the beginning of clearly defined periodicity in the waveform and at the appearance of low frequency energy in the spectrogram.

- **End:** when periodicity terminates or at closure onset; if before a lateral, when formant transition approaches steady state; if before a nasal, at the point where formants show a step-wise shift and when intensity shows a steep decline.

### §2.2.2 Tokens with positive VOTs

Stimuli with positive VOTs were created using the "progressive cutback and replacement method" (version 31) (Winn, 2020) implemented in Praat (Boersma & Weenink, 2022). The process takes a voiced token and progressively deletes its vowel onset and replaces it with an approximately equivalent amount of onset taken from the word's voiceless minimal-pair counterpart. Winn's Praat script provides a GUI that greatly simplifies the generation of highly natural sounding stimuli.

For each minimal pair a continuum of 31 tokens was generated between 0ms and 130ms with a step-size of 5ms. A token of the voiced category from each pair was selected to be the base sound file to make the continuum. All four minimal pair continua were created using the same aspiration sound which was excised from one of the voiceless tokens produced by the talker.

We set the fundamental frequency (F0) to covary with VOT according to the natural correlation found in the measurements of our test talker's recorded minimal pair tokens. We first ran a linear regression that predicted the talker's F0 values from the measured VOT values. This gave the expected F0 value at every 5 ms interval of VOT. To produce the expected F0 values as the VOT tokens were generated, the start-point F0 value (where $VOT = 0ms$ and $F0 = 246Hz$) and the end-point F0 value (where $VOT = 150ms$ and $F0 = 252Hz$) were entered into the Praat script. The resulting F0 values for each token were not identical across the minimal pair words as shown in Figure **??** (the f0 measurements obtained over the first 5ms from vowel onset) but were sufficient for our aim to keep F0 positively correlated with VOT.

The vowel cut-back ratio was set at 0.33 which translates into 0.33 ms vowel reduction for every 1ms of additional VOT. This ratio followed the estimated vowel duration-VOT trade-off for dip-tip minimal pair tokens reported in Allen and Miller (1999). The maximum vowel cut-back allowed was 0.5ms to avoid the short vowel in "dip" becoming too short.

Aspiration intensity was allowed to covary positively with VOT according to the script's default setting. The script manipulates intensity by attenuating the original aspiration sound by a user-specified value. For a continuum with $n$ steps with the value set to 6 dB (the default) the aspiration sound will be lowered by 6 dB for step 1 (the voiced end of the continuum) and then gradually interpolated to increase in intensity across the continuum so that there will be no attenuation for the $n$th step. Attenuation of the aspiration continua in this experiment was made over 21 steps (0-100ms) which resulted in a range of 55.5 dB to 60.5 dB. Aspiration for stimuli with over 100 ms VOT was maintained at 60.5 dB.

#### §2.2.3 Tokens with pre-voicing ('negative VOTs')

Stimuli with pre-voicing required a separate approach because Winn's Praat script does not yet support the creation of tokens with pre-voicing that are natural sounding.[S1] Pre-voicing stimuli were created by prepending pre-voicing generated from naturally produced tokens (described below) that were edited with a separate process.

Pre-voicing in 5ms increments were generated from a a clear pre-voicing waveform excised from a pre-voiced token produced by the talker. To achieve the desired duration a duration factor was first computed and then converted with the "lengthen (overlap-add)" function in Praat based on the PSOLA algorithm. For example, if the desired amount of pre-voicing was 50ms then the duration factor would be 50ms/length of the original pre-voicing sample. Each pre-voicing step was then pre-pended to a token with 0ms VOT. Each of these 0ms tokens was generated with the same aforementioned progressive-cutback-and-replacement Praat script individually. Each 0 ms VOT token would have an F0 value that corresponded to the amount of pre-voicing duration based on the predictions of the linear model. No vowel-cut back was implemented for pre-voiced tokens.

[S1] It can however, produce pre-voicing sufficiently well for demonstration purposes (see video demo at https://www.youtube.com/watch?v=-QaQCsyKQyo).

### §2.2.4 Annotation of resynthesized stimuli (used in visualizations and for ideal observer/adaptor analyses)

All resynthesized stimuli were annotated for VOT, pre-voicing, vowel duration, and f0 following the same procedure as in the original recordings. We follow previous work on speech perception and combined the pre-voicing and VOT data, treating pre-voicing as negative VOT. This simplifies plotting and analysis. It does, however, make assumptions that we revisit in the general discussion.

For reasons we describe next, we then corrected and simplified the measurements derived from these annotations. We used these corrected measurements for all visualizations of the stimuli (e.g., in Figure 6 in the main text) and in our ideal observer and adaptor analyses. In the data shared on OSF, the corrected cue values are stored as Item.CUE_NAME (e.g., Item.VOT) and the measured values are stored as Item.CUE_NAME.measured (e.g., Item.VOT.measured).

We used the *corrected* rather than *measured* f0 and vowel duration values. This removes measurement noise that might be introduced by the annotation procedure, and makes f0 and vowel duration values more akin to our VOT values. Specifically, the f0 values were linearly interpolated between the f0 values of the voiced and voiceless tokens of the minimal pair. For vowel durations, we used the vowel durations of the *dip-tip* pair, which sidesteps the difficulty of annotating the boundary between the vowel and the coda for *din-tin* and *dill-till* word pairs. This assumes that listeners perceive the same vowel duration for the different words that have the same VOT duration. This assumption is likely a reasonable approximation given that the stimuli were generated from word recordings that were made in the same context and spoken by the same individual in a similar speech style and rate. Figure S1 shows uncorrected and corrected measurements of all stimuli while Table S1 shows their correlation values after correction.

Table S1
*Cue correlations after measurement corrections*

| Cues | VOT | f0 | Vowel duration |
|---|---|---|---|
| VOT | 1.00 | 0.98 | -0.97 |
| f0 | 0.98 | 1.00 | -0.95 |

| Cues | VOT | f0 | Vowel duration |
|---|---|---|---|
| Vowel duration | -0.97 | -0.95 | 1.00 |

# §3 Additional information on participant exclusions

Beyond the participants mentioned in the main text, an additional 115 participants previewed the front page of the experiment but did not start or complete it. Our technical setup did not allow us to distinguish between participants who started the experiment and did not complete it, and participants who only previewed the experiment. We suspect though that the clear majority of the 115 participants falls into the latter category: crowdsourcing participants often preview multiple experiments and then select one to complete. Unlike in lab-based experiments, for which participants' right to stop the experiment at any point can be costly—both in terms of effort and perceived social cost—exercising this right in web-based experiments is essentially cost free. One appealing feature of the Prolific platform is that it records these numbers (unlike some other platforms). We report them here in the hope that this will become standard. Viewed across different studies, these data will have the potential to identify unintended biases in participant recruiting.

Next, we provide additional information on participants' performance during catch trials and on labelled trials (for which there is a clearly correct response). Both of these measures were used to exclude participants.

## §3.1 Performance on catch and labelled trials

Of the 122 participants that completed the experiment, 1 committed more than 3 errors ($< 84\%$ accuracy). Labelled trials provided another measure to check participant attention during the experiment. Performance on labelled trials was good across all exposure conditions with participants in the +40 condition committing the most errors. This is not unexpected as the +40 condition had the furthest shifted with the most unexpected cue-category mappings.

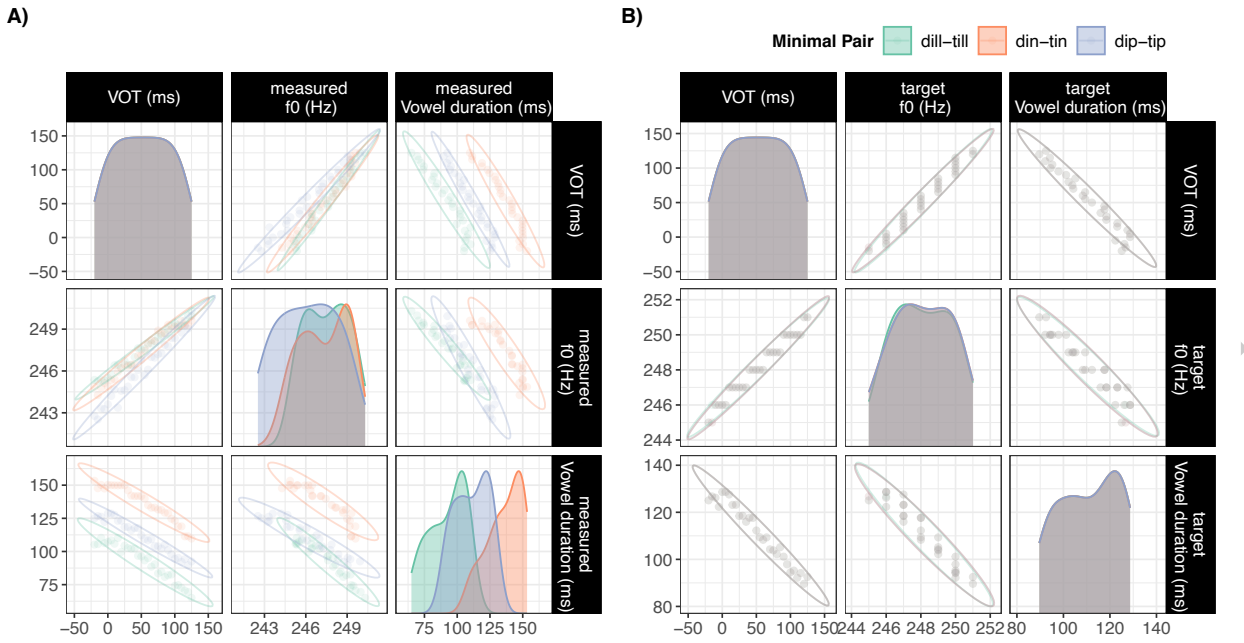


*Figure S1.* Correlations of VOT, f0 and vowel duration of the synthesized stimuli **A)** before and **B)** after measurement corrections.

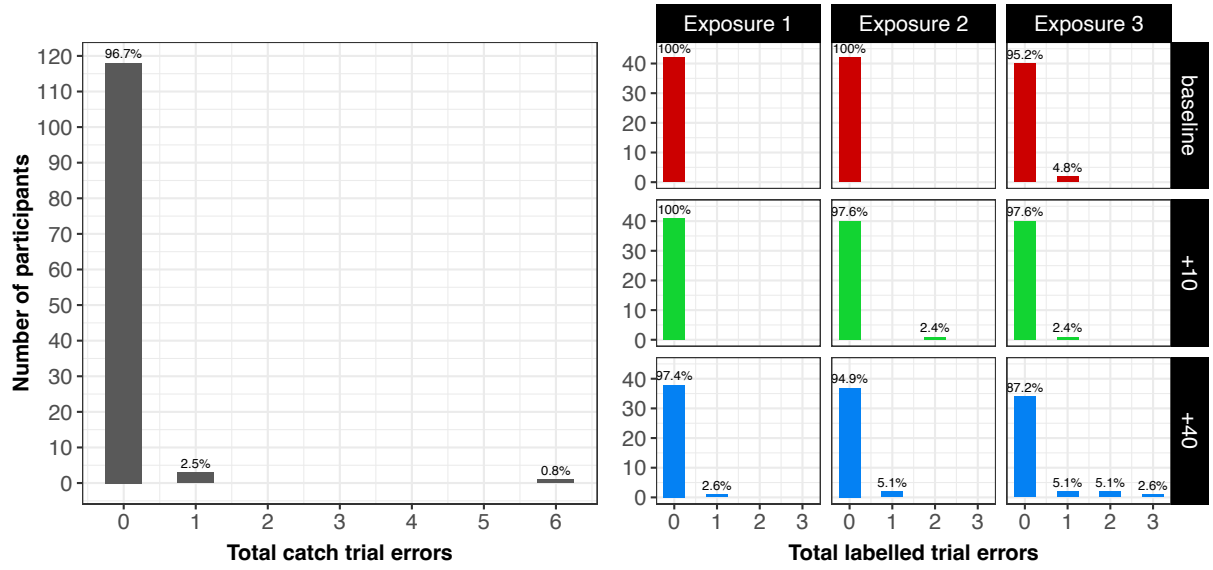


*Figure S2.* Total errors committed on catch trials and labelled trials. Percentage values indicate the proportion of participants in the respective groups.

# §4 Obtaining predictions of idealized listeners

In the main text, we employ ideal observers (specifically, the ASP framework described in Xie et al., 2023) to relate our results to those expected from idealized listeners. This includes predictions based on a *'typical' listener's prior experience* and predictions based on *idealized learners for each exposure condition* that have fully learned the exposure distributions. The former provides a baseline against which to compare listeners' behavior at the start of the experiment. The latter provides a baseline against which to compare changes in listeners' behavior as a function of exposure. In this section, we describe how we derived those predictions.

## §4.1 Idealized pre-exposure listeners

Estimates of listeners' prior expectations about the realization of phonetic categories require at least three ingredients: (1) a database that can be reasonably assumed to approximate the types of previous experiences that listeners in the experiment would draw on to process the stimuli in the experiment; (2) a model that is used to approximate the computational consequences of listeners' representations derived from previously experienced speech stimuli; and (3) a procedure to estimate / parameterize that model based on the database (see also Feldman et al., 2009; Kronrod, Coppess, & Feldman, 2016; Norris & McQueen, 2008; Persson & Jaeger, 2023; Tan et al., 2021; for a helpful review, see Schertz & Clare, 2020). In this section, we describe the decisions we made regarding this three components.

### §4.1.1 A database of L1-US English word-initial /d/ and /t/ productions

We considered two databases to approximate listeners' prior expectations about the phonetic realization of /d/ and /t/ by a female L1 talkers of US English (as the one employed in our experiment). The two databases contain productions of word-initial /d/ and /t/ in isolated word productions and in connected read speech, respectively (Chodroff & Wilson, 2017, available on OSF at https://osf.io/k6djr/).[S2] The connected speech database includes word-initial stop

[S2] We thank Eleanor Chodroff for adding the isolated speech data to OSF, and for prompt and helpful responses to our questions.

production by 180 adult L1-US English talkers (102 female), while the isolated speech database was made up of recordings of isolated utterances of stop-initial CVC syllables by 24 L1-US English talkers (13 female, see Chodroff & Wilson, 2017, pp. 33, 37–39 for details).

We filtered both databases to only the /d/ and /t/ tokens. For both databases we kept only female talkers to match the gender of our test talker. We kept talkers that had at least 15 tokens of each category. We removed tokens with anomalous f0 measurements, including tokens that were likely due to pitch-halving (identified by examining individual pitch plots) as well as tokens with f0 measurements below 150 Hz (as these are likely due to measurement error; Eleanor Chodroff, p.c.). Of the remaining talkers we ensured that an equal number of /d/ and /t/ tokens within each talker were sampled. The sample size was determined by first counting the number of tokens available for each talker and category, and then taking the lower of the two.

This left a total of 5,756 tokens from 92 female talkers: 4,704 tokens by 79 female talkers from the connected speech database and 1,052 tokens by 13 female talkers from the isolated speech database. This is the data shown in Figure 6.

For the predictions for idealized pre-exposure listeners, we decided to focus on the isolated speech data. While this database is substantially smaller (13 female talkers), the conditions under which the recordings were elicited more closely resemble the conditions used to record our stimuli from which the test stimuli were created. Both recordings were sampled from a similar population pool, i.e. undergraduate native L1-US English speakers and both were recordings of isolated utterances. Nonetheless there were differences in the way talkers were instructed to produce the tokens that may have contributed to differences we see between our stimuli and the database's. The isolated database tokens were CV*t* syllables (e.g., *dot*, *tot*) produced in the context of a carrier phrase "*Say* _____ *again.*". The participants were told to speak at a normal rate and to make a slight pause after "Say" and before "again" (Chodroff & Wilson, 2014). The recordings of our talker were made without a carrier phrase.

#### §4.1.2 A model of listeners representations of /d/ and /t/ categories

We used the adaptive speech perception (ASP) framework described in Xie et al. (2023) as a convenient way to approximate how listeners map the acoustic-phonetic properties of the speech

input onto phonetic categories like /d/ and /t/. Following Xie and colleagues, we assume that category representations can be approximated as multivariate Gaussian distributions in the phonetic space. Specifically, we describe /d/ and /t/ as multivariate Gaussian categories of VOT (in ms), f0 (in Mel), and vowel duration (in ms). As mentioned in the main text, we follow previous work and treat pre-voicing as negative VOT, and revisit the consequences of that decision in the general discussion. The choice of multivariate Gaussian categories strikes a compromise between substantially less parsimonious models (e.g., exemplar models, neural networks, LDA) and even more parsimonious models (e.g., independent Gaussians for each phonetic cue, as discussed in more detail in Xie et al., 2023). Also following Xie and colleagues, we included perceptual noise in the ideal observer models. Perceptual noise was assumed to be independent for each phonetic cue. Noise estimates for VOT and f0 were obtained from previous work (Kronrod et al., 2016, $sigma_{noise} = 80ms^2$ and $878Mel^2$, respectively), the perceptual noise for vowel duration was set to the same value as for VOT given that both are durational cues ($sigma_{noise} = 80ms^2$).

#### §4.1.3 A procedure to fit the parameters of the model

To train ideal observers (or other models of phonetic representations), it is necessary to make assumptions about listeners' representations of phonetic categories. Previous work has often derived estimates under the implicit assumption that listener learn and maintain a single, *talker-independent*, representation for each phonetic category across talkers. This assumption typically has taken one of two forms. One approach is to entirely ignore talker identity during the estimation of phonetic categories. This is illustrated in Figure S3 which shows the distribution of /d/ and /t/ over VOT, f0, and vowel duration in isolated (top left) and connected (bottom left) speech. Ellipses indicate the phonetic category representations that would be estimated from these data. Another approach is to first normalize each cue within each talker—e.g., by subtracting the talker's cue mean from each token (e.g., McMurray, Cole, & Munson, 2011; McMurray & Jongman, 2011; for review, see Apfelbaum & McMurray, 2015; Weatherholtz & Jaeger, 2016). Phonetic categories are then estimated over these normalized cues. Under this approach, cues are interpreted in a talker-dependent way but the phonetic category representations remain

talker-independent. The resulting representations are shown in Figure S3C and S3D.

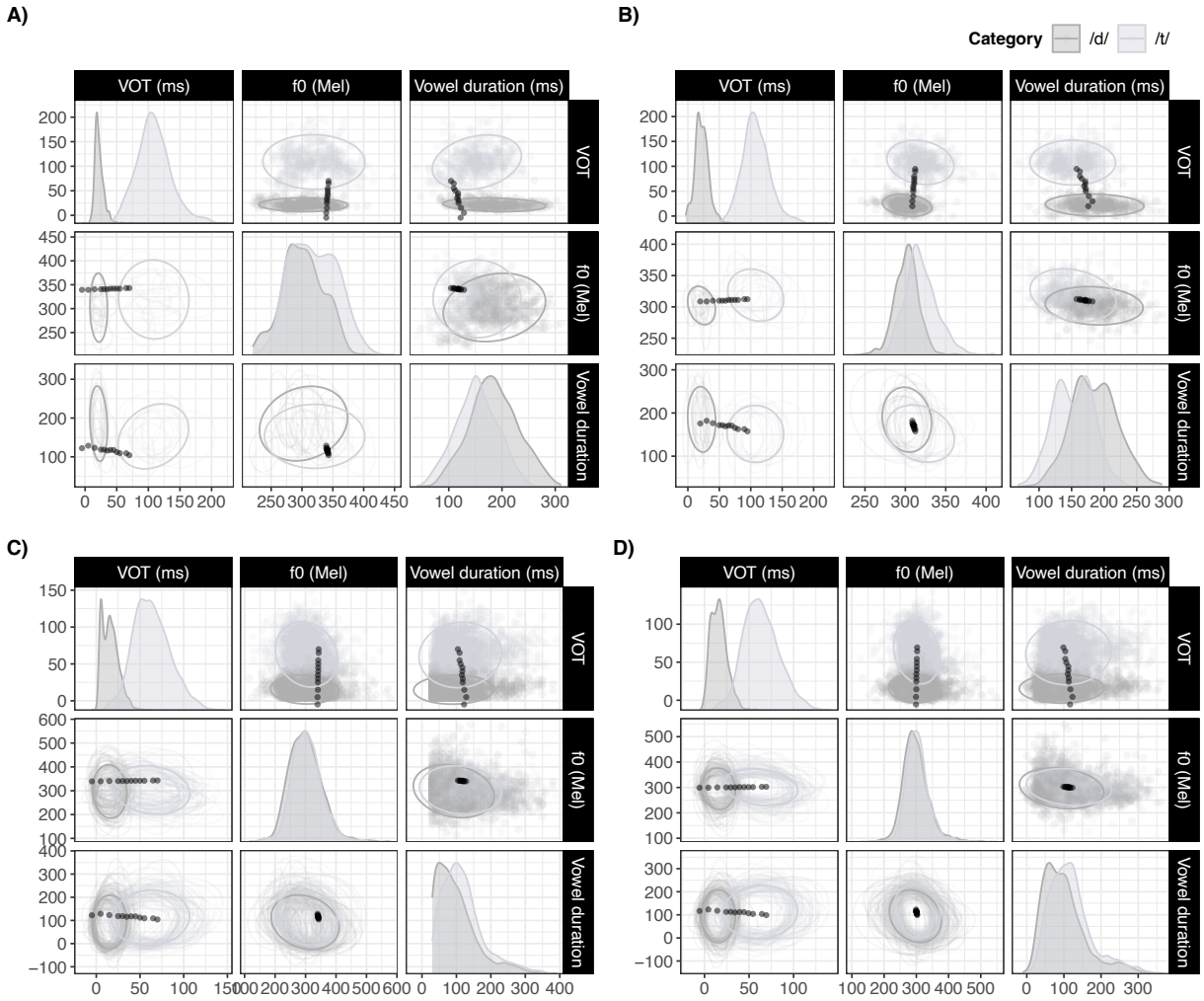


*Figure S3.* Placement of test tokens of our experiment relative to distribution of three important cues to word-initial stop-voicing in L1-US English. **Panel A:** Isolated speech of initial /d/ and /t/ syllables by 13 female talkers. **Panel B:** Same as **A** but talker-normalized. **Panel C:** Connected speech of word-initial /d/ and /t/ by 79 female talkers. **Panel D:** Same as **C** but talker-normalized. Following Xie et al. (2023), we added the overall cue mean to all talker-normalized tokens in order to show normalized tokens in the same space as unnormalized tokens (thus keeping the phonetic space constant across panels).

The two approaches to talker-independent category representations have different trade-offs. When cues are not talker-normalized, the resulting category representations inherit not only within-category variability but also variability across talker, over-estimating the actual category covariances for any given talker. Normalizing cues addresses this problem. It does, however, remove potentially useful information about the correlation between cues from the signal:

compared to unnormalized cues in Panel A, the talker-normalized cues in Panel B contain less information about the covariance of VOT, f0, and vowel duration.

The two approaches also share potentially important shortcomings. Specifically, neither approach is well-suited if the correlation between cues *within* talkers differs from the correlation between talker means of those cues. The /d/ and /t/ categories in Panel A conflate the two source of covariance; the categories in Panel B completely removed any information about covariance in talker means. This is problematic, as there is strong evidence that such correlations exists (Chodroff & Wilson, 2017, 2018; Sonderegger, Stuart-Smith, Knowles, Macdonald, & Rathcke, 2020; Theodore, Miller, & DeSteno, 2009), and that listeners have strong expectations about, at least some types of, correlations between talker means (Idemaru & Holt, 2011, 2020; Schertz et al., 2016; **OTHERs?**).[S3] For the same reasons, neither of the two approaches considered above captures correlations between talkers' category means and category (co)variances. Both approaches would, for example, miss if category variance generally increases for larger category means (or vice versa).

An alternative to the two approaches described so far is to derive multiple *talker-dependent category representations* (as assumed in e.g., Kleinschmidt & Jaeger, 2015). Under this approach, expectations for a 'typical' talker are then derived by averaging over the talker-dependent category representations. And expectations for a typical talker of a certain type (e.g., a female talker of a certain age) are derived by averaging over the talker-dependent category representations learned from previous talkers of that type. In the only study we are aware of that has directly compared talker-independent and talker-dependent models of listeners' prior expectations, talker-dependent representations seemed to better describe human behavior (Xie, Buxó-Lugo, et al., 2021), though it is important to keep in mind that this study focused on a single supra-segmental contrast (question vs. statement prosody).

Given the uncertainty about the most adequate approach, we implemented two of the three alternatives for the present study. For most of the main text, we rely on talker-independent category representations over the *un*normalized distribution of VOT, f0, and vowel duration

[S3] Additionally, the category means for /d/ and /t/ (solid points) for unnormalized cues are simply the weighted average of the talkers available in the data. If talkers contribute different amounts of data to the database, this means that the category means might disproportionately depend on a subset of the data.

(Panel A of Figure S3). To this end, we obtained the mean and covariance matrices for /d/ and /t/ across all talkers in that unnormalized cue space. This talker-independent model is used to create the predictions for an idealized pre-exposure listeners in, for example, Figure 5C. Given the small number of talkers in the isolated speech database (13), its restriction to University students from a single campus, and the stimuli being limited to selected phonetic contexts this result is unlikely to be representative of a typical listener's exposure to everyday speech. To safeguard against providing a false sense of strength given by this point estimate, we divided the database into five folds that were randomly sampled within all 26 unique combinations of Talker and category. We then trained five separate ideal observers—one for each of the five folds. This gives us the 95% CIs shown in Figure 5C.[S4]

In the general discussion, however, we present a talker-dependent model to the data from all five folds. For this, we estimated the category means and covariances separately for each talker. This gave us the range of talker-specific PSE estimates shown on the righthand-side of Figure 8.

## §4.2 Idealized learners that have fully learned the exposure distribution

To construct idealized learners that fully learned the exposure distribution, we followed a similar approach as for the idealized listeners. We again used ideal observers to approximate such a learner's category representations. Unlike for the idealized listeners described above, the idealized learners only consider the only cue that was manipulated in the experiment (VOT). This decision was made since we constructed our stimuli such that the other two cues (f0 and vowel duration) were perfectly correlated with VOT, not providing any additional information. For each exposure condition we constructed ideal observers to simulate a learner that fully learned the exposure distribution (i.e. the cumulative distribution after having experienced all 144 trials). Such an idealized learner would therefore categorize the stimuli at test based purely on the statistics of exposure input.

[S4] An alternative would have been to randomly distribute talkers into the folds. With few talkers in the database this would have meant that each prior PSE estimate would have been based on data from only 2-3 talkers. Since the aim is to simulate an adult L1 US English listener's knowledge the prior should ideally be estimated on as many talkers as possible. The first option strikes the best compromise given the small database we had at hand.

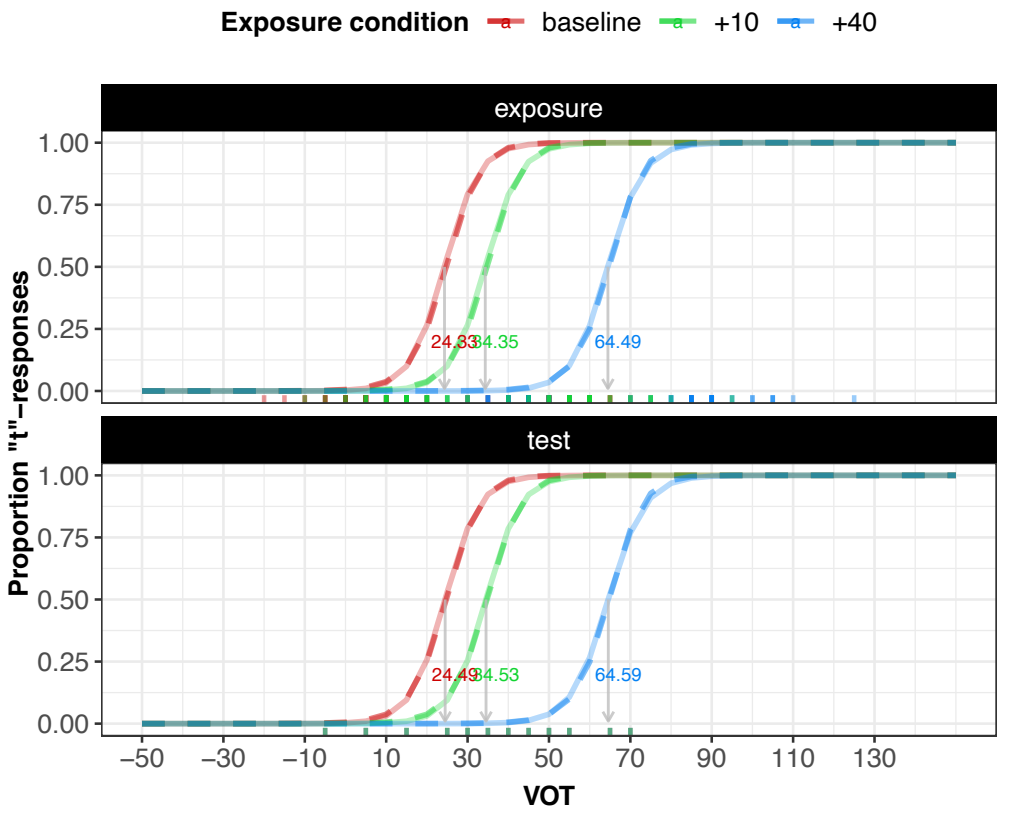


*Figure S4.* The expected categorization functions predicted by a logistic regression fit to the responses of an idealized learner by exposure condition and phase (solid lines). The idealizes learner's proportion of "t"-responses (dashed lines) converge with the logistic regression fit along the assessed VOT region and are therefore partially obscured by the logistic regression fitted lines. Grey arrows point to expected PSEs of the logistic regression fit.

### §4.3 Putting models and listeners on the same scale: equating potential biases in the estimation of intercepts, slopes, and PSEs

Figure 5C in the main text shows PSE estimates for the participants in our experiments. These estimates are based on the perceptual model contained in the Bayesian psychometric mixed-effects model we used to analyze participants' responses, and thus reflect PSEs that are corrected for the rate of attentional lapses. Of note, this perceptual model assumes linear effects of VOT on the log-odds of "t"-responses. We made this assumptions for the sake of simplicity, and in order to avoid over-fitting. On the one hand, auxiliary analyses presented in section §5.9 replicate all tendencies reported in the main text while relaxing the linearity assumption. On the other hand, it is possible that the linearity assumption introduced biases into the estimation of

participants' intercepts, slopes, and PSEs. Critically, the ideal observers that we use to describe idealized listeners before and after exposure predict non-linear effects of VOT [since /d/ and /t/ do *not* have equal variances; for details, see (Bicknell, Bushong, Tanenhaus, & Jaeger, under review; Kleinschmidt & Jaeger, 2015; Kronrod et al., 2016). Thus, instead of directly calculating predicted intercepts, slopes, and PSEs from the ideal observers, we estimated their intercepts, slopes, and PSEs paralleling the analysis approach for the human data. This makes sure that any biases in the estimation of human intercepts, slopes, and PSEs that are introduced by our analysis approach are also taken into account for the idealized listeners and learners.

Specifically, we applied the following procedure to each of the five idealized observers representing idealized listeners prior to exposure (one for each cross-validation fold), and each of the three ideal observers representing idealized learners (one for the three exposure condition). We first sampled 1e+12posterior responses from the ideal observer at each of the VOT steps. For example, for test blocks, we sampled 1e+12 responses from the model at each of the 12 VOT steps. These responses were sampled while assuming a lapse rate of 0 and a uniform response bias. We then fit a logistic regression model—predicting "t"-responses—to the sampled responses, with VOT as the predictor and the sampled responses as the outcome. By emulating a zero lapse rate and using an ordinary logistic regression, we derive estimates of intercepts, slopes, and PSEs for each ideal observer that are directly comparable to the estimates for human listeners (which reflect the *lapse-corrected* PSE of the perceptual model component of the psychometric mixed-effects model fit to listeners' responses).

Figure 5C in the main text presents prediction lines that were estimated based on the test blocks. This decision was made since procedure described here produced very similar bias-corrected predictions for exposure and test blocks, and we felt that introducing all complexity described here in the main text would distract from the main message of the paper. Figure S5 shows a version of Figure 5, in which the prediction lines and ribbons have been adjusted separately for each block. In this variant of the figure, we also include an additional panel that shows the block-by-block changes in the *intercepts* and *slopes* of listeners' categorization functions, as well as the ideal observers' predictions for these parameters (discussed further in sections §5.4 and §5.5).

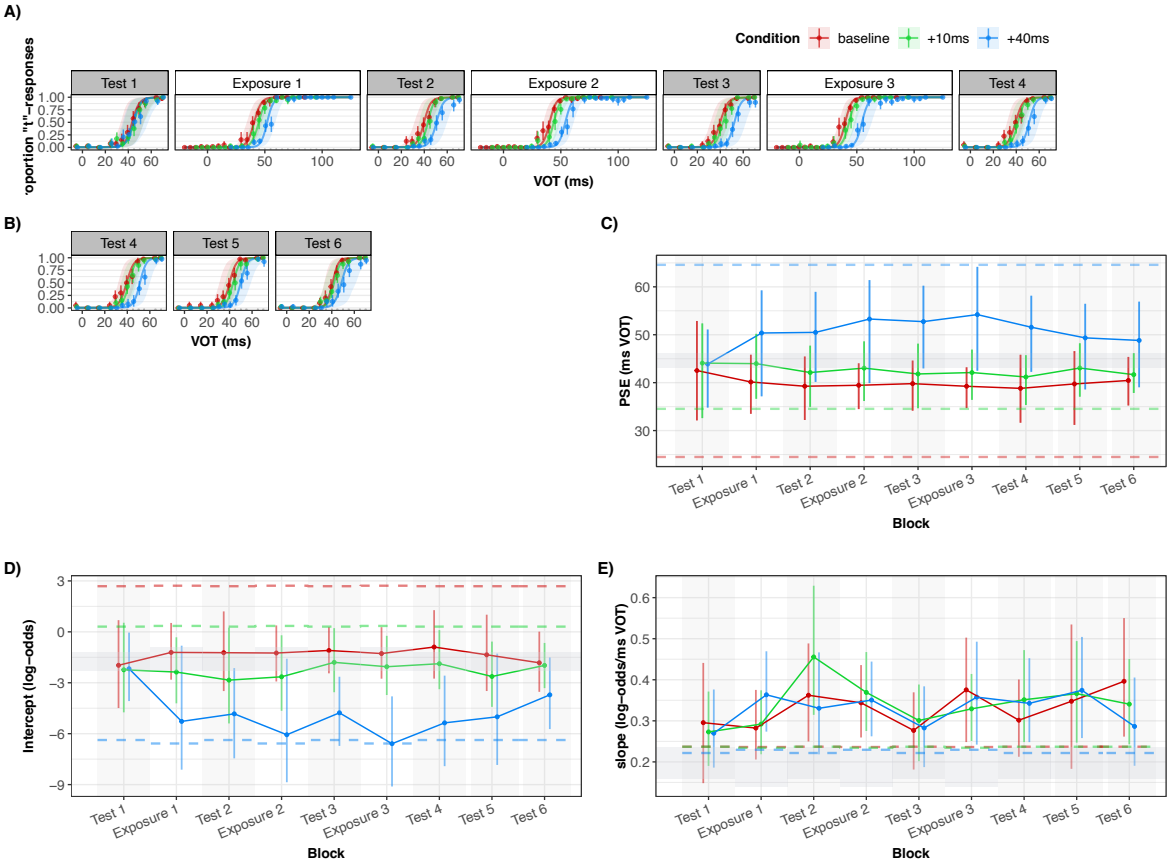


*Figure S5*. Same as Figure 5 but also showing block-to-block changes in the *intercepts* and *slopes* of listeners' categorization function (PSE = -intercept / slope). Additionally, the prediction lines for idealized learners are adjusted on a block-by-block basis for potential biases in the estimation of intercepts, slopes, and PSEs that might be introduced by the linearity assumption of the psychometric mixed-effects model (that effects of VOT on listeners' responses are linear in log-odds of responding "t").

# §5 The Bayesian psychometric mixed-effects model: additional information and hypothesis tests

We first present some additional information about the psychometric mixed-effects model used for the analysis of participants' categorization responses. Then we present detailed summary tables of the Bayesian hypothesis tests about listeners' *PSE* for the test and exposure blocks, followed by parallel summary tables for the *slope* of listeners' categorization function. Then, we present an auxiliary analysis that assess changes in lapse rates across the exposure and test blocks (the main analyses assumed constant lapse rates within each type of block). These analyses validate our

decision to assume constant lapse rates within each type of block. Finally, we present analyses that combined the exposure and test data, while relaxing the linearity assumption for the effects of VOT. All analyses presented in this section are variants of the same psychometric mixed-effects model, which we introduce next.

## §5.1 Additional information about the Bayesian psychometric mixed-effects model

We analyzed participants' categorization responses during exposure and test blocks in two separate Bayesian mixed-effects psychometric model (Kuss et al., 2005; Prins, 2011, 2019b; Schütt, Harmeling, Macke, & Wichmann, 2015). The psychometric model is an extension of mixed-effects logistic regression that also takes into account attentional lapses. The mixed-effects psychometric model describes the probability of "t"-responses as a weighted mixture of a perceptual and a lapsing model. The perceptual model predicts responses on trials where participants pay attention and respond based on the stimulus. We implemented the perceptual model as a mixed-effects logistic regression, predicting "t"-responses from exposure condition (backward difference coded, comparing the +10ms against the baseline condition, and the +40ms against the +10ms shift condition), test block (backward difference coded from the first to last test block), VOT (Gelman scaled), and their full factorial interaction. The model included by-participant random intercepts and slopes for all within-participant manipulations (block and VOT) and by-item random intercepts and slopes for all within-participant manipulations (exposure condition, block, VOT).

We fit the two psychometric models for exposure and test blocks using the package `brms` (Bürkner, 2017) in R (R Core Team, 2022; RStudio Team, 2020).[S5] Predictor coding and priors were identical across both models. To facilitate comparison of effect sizes across predictors, we

[S5] Here and throughout the text, we refer to one model for the test blocks and one model for the exposure blocks. However, more specifically, we fit several variants of each model that only differed in the way that predictors were nested within the perceptual model. In addition to the standard formulation of the perceptual model, which expressed the effect the effects of exposure condition, block, and VOT as the full factorial interactions (`response ~ condition * block * VOT`), we also fit variants that yielded the simple effects of block within each condition (`response ~ condition / (block * VOT)`), the simple effects of condition within each block (`response ~ block / (condition * VOT)`), or separate intercept and VOT slope estimates for each exposure condition and block (`response ~ 0 + (condition * block) / VOT`). All of these variants are prediction-equivalent, but each of them facilitate the formulation of different hypothesis tests of relevance to our questions. For details, we refer to the R markdown document that this SI is generated from. The R markdown contains the R code used to specify the different models and hypothesis tests.

standardized continuous predictors (VOT) by dividing through twice their standard deviation (Gelman, 2008). We also centered VOT based on the mean of the *test* blocks. This makes sure that all other effects—e.g., the effects of exposure condition and block—are analyzed at the same VOT across the two separate models. Following previous work from our lab (Hörberg & Jaeger, 2021; Xie, Liu, et al., 2021) we used weakly regularizing priors to facilitate model convergence. For fixed effect parameters, we used Student priors centered around zero with a scale of 2.5 units (following Gelman, Jakulin, Pittau, & Su, 2008) and 3 degrees of freedom, with the exception of VOT which had a wider scale of 15 units because of convergence issues (see §5.3 for elaboration). For the lapse rate we used the `brms` default logistic prior centered around 0 with a scale of 1 unit. This assumes uniformity for $0 < lapserates < 1$ while down-weighting the extreme values. For random effect standard deviations, we used a Cauchy prior with location 0 and scale 2, and for random effect correlations, we used an uninformative LKJ-Correlation prior with its only parameter set to 1, describing a uniform prior over correlation matrices (Lewandowski, Kurowicka, & Joe, 2009). Four chains with 2000 warm-up samples and 2000 posterior samples each were fit. No divergent transitions after warm-up were observed, and all $1 < \hat{R} < 1.01$.

## §5.2 PSE results for test blocks

The main text contains summary tables for the simple effects of exposure condition within each test block (Table 2) and the simple effects of block within each exposure condition (Table 3). The main text also refers to differences in the rate of block to block changes across exposure conditions. These differences correspond to the interactions between exposure conditions and block in the psychometric mixed-effect model. These interactions are summarized in Table S2.

### §5.2.1 Differences in the rate of change between exposure conditions and block

## §5.3 PSE results for exposure blocks

While the standard formulation of the psychometric mixed-effects model converged (Tables S5 and S11) the nested formulations from which we obtain simple effects, returned several divergent transitions during sampling. Divergent transitions signal the need for caution in model

Table S2
*Did the rate of block-to-block changes in PSEs differ across exposure conditions? This table summarizes the interactions between exposure condition and block—specifically, whether the differences between exposure conditions changed from test block to test block.*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Difference in +10 vs. baseline** | | | | | |
| Block 1 to 2: increased $\Delta_{PSE}$ | -0.85 | 0.78 | [-2.17, 0.63] | 5.4 | 0.84 |
| Block 2 to 3: increased $\Delta_{PSE}$ | 0.34 | 0.77 | [-1.14, 1.76] | 0.5 | 0.32 |
| Block 3 to 4: increased $\Delta_{PSE}$ | 0.06 | 0.77 | [-1.38, 1.53] | 0.9 | 0.47 |
| *Block 1 to 4: increased* $\Delta_{PSE}$ | -0.42 | 1.26 | [-2.76, 1.96] | 1.7 | 0.63 |
| Block 4 to 5: decreased $\Delta_{PSE}$ | -0.33 | 0.60 | [-1.43, 0.79] | 0.4 | 0.29 |
| Block 5 to 6: decreased $\Delta_{PSE}$ | 1.03 | 0.65 | [-0.23, 2.16] | 11.9 | 0.92 |
| *Block 4 to 6: decreased* $\Delta_{PSE}$ | 0.70 | 0.82 | [-0.9, 2.18] | 3.8 | 0.79 |
| **Difference in +40 vs. +10** | | | | | |
| Block 1 to 2: increased $\Delta_{PSE}$ | -2.36 | 0.89 | [-3.81, -0.75] | 57.8 | 0.98 |
| Block 2 to 3: increased $\Delta_{PSE}$ | -1.16 | 0.83 | [-2.59, 0.31] | 10.0 | 0.91 |
| Block 3 to 4: increased $\Delta_{PSE}$ | -0.27 | 0.82 | [-1.69, 1.16] | 1.7 | 0.63 |
| *Block 1 to 4: increased* $\Delta_{PSE}$ | -3.78 | 1.22 | [-5.87, -1.45] | 84.1 | 0.99 |
| Block 4 to 5: decreased $\Delta_{PSE}$ | 1.14 | 0.77 | [-0.24, 2.51] | 11.4 | 0.92 |
| Block 5 to 6: decreased $\Delta_{PSE}$ | 0.45 | 0.77 | [-0.98, 1.79] | 2.6 | 0.72 |
| *Block 4 to 6: decreased* $\Delta_{PSE}$ | 1.59 | 0.99 | [-0.3, 3.32] | 12.7 | 0.93 |
| **Difference in +40 vs. baseline** | | | | | |
| Block 1 to 2: increased $\Delta_{PSE}$ | -3.16 | 1.02 | [-4.96, -1.18] | 79.0 | 0.99 |
| Block 2 to 3: increased $\Delta_{PSE}$ | -0.82 | 1.08 | [-2.75, 1.14] | 3.4 | 0.77 |
| Block 3 to 4: increased $\Delta_{PSE}$ | -0.20 | 1.08 | [-2.15, 1.74] | 1.3 | 0.57 |
| *Block 1 to 4: increased* $\Delta_{PSE}$ | -4.19 | 1.71 | [-7.22, -0.93] | 45.8 | 0.98 |
| Block 4 to 5: decreased $\Delta_{PSE}$ | 0.80 | 0.92 | [-0.97, 2.49] | 4.2 | 0.81 |
| Block 5 to 6: decreased $\Delta_{PSE}$ | 1.48 | 0.94 | [-0.36, 3.12] | 10.9 | 0.92 |
| *Block 4 to 6: decreased* $\Delta_{PSE}$ | 2.27 | 1.27 | [-0.12, 4.44] | 16.5 | 0.94 |



interpretation because it indicates unreliability of the posterior estimate (see https://mc-stan.org/misc/warnings.html#runtime-warnings for explanation of warning types).

A possible solution to eliminating divergent transitions in this model would be to relax the variance of the prior distribution assumed for the slope parameter (*VOT_gs* in the model formula). Indeed, in earlier fits of the test phase data we faced convergence issues when we used the recommended weakly regularising prior (student-t distribution with SD = 2.5; Gelman et al. (2008)) for the slope parameter. Diagnostic plots revealed a slight bi-modality in the posterior distribution which may have been an impediment to smooth sampling; this was solved by increasing the prior distribution's variance to 15 units which allowed the inclusion of data points

that were further from the mean.[S6] However, fitting the nested exposure models here with yet more accommodating priors would detract from having an identical set of priors across all models fitted for test and exposure. Given that both the simple effects models returned only a few divergent transitions (3 in the nested block and 9 in the nested condition model) and showed fairly good $\hat{R}$ values ($1 < \hat{R} < 1.1$) and ESS values we opted not to explore further solutions for them at this stage.

Consistent with the trend observed in test blocks, categorization boundaries were clearly separated between the conditions in all exposure blocks and grew progressively larger with more exposure (Table S3). Evidence for this effect was stronger in the +40 vs. +10 comparison ($11 \leq$ BFs $\leq 257$) with the difference in PSE reaching its maximum in the final exposure block ($\hat{\beta} = -3.99$, 90%-CI = $[-5.61, -2.088]$, $BF = 91$, $p_{posterior} = 0.989$). The difference between +10 and baseline widened in exposure block 2 ($\hat{\beta} = -1.4$, 90%-CI = $[-2.57, -0.211]$, $BF = 29$, $p_{posterior} = 0.967$) but narrowed in the final block ($\hat{\beta} = -0.98$, 90%-CI = $[-2.147, 0.2]$, $BF = 11.4$, $p_{posterior} = 0.919$). The smaller difference in PSE between the +10 and baseline conditions reflected the pattern in test blocks.

Analysis of boundary shifts between exposure blocks show a similar incremental pattern albeit with weaker evidential support ($0.9 \leq$ BFs $\leq 3$ Table S4). In the +10 condition, categorization functions shifted leftwards with greater exposure while in the +40 condition categorization moved further rightwards. Categorization in the baseline condition showed an overall smaller shift between exposure blocks 1 to 3.

The effects of VOT (slope) did not change significantly between conditions within each exposure block (Table S9) and between blocks within exposure condition (Table S10). This conforms to the trend found in slope analyses for test blocks as well as the predictions of the ideal observers. Table S5 and Table S11 report the effects of interactions between exposure condition and block (PSE); and exposure, condition, block and VOT respectively (slope). With respect to PSE changes, the difference between +10 and baseline held constant from exposure block 1 to exposure block 2 and increased marginally from exposure block 2 to exposure block 3 ($0.061 \leq$

[S6] A discussion on the theoretical and practical consequences of prior-setting approaches can be found in (Gelman, Simpson, & Betancourt, 2017).

estimates $\leq 0.35$; $0.9 \leq$ BFs $\leq 0.35$). Between the +40 and +10 conditions, PSE differences increased with more exposure as did PSE differences between +40 and the baseline condition. This trend reflects that of test blocks but evidential support was weaker than that found in test block comparisons ($2.1 \leq$ BFs $\leq 1.8$).

### §5.3.1 Simple effects of condition within each exposure block

Table S3
*This table summarizes the simple effects of the exposure conditions for each exposure block. Note that rightward shifts of the categorization function (and its PSE) correspond to negative estimates (lower intercepts in predicting the log-odds of "t"-responses).*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Exposure block 1** | | | | | |
| +10 vs. baseline $< 0$ | -1.25 | 0.65 | [-2.4, -0.09] | 23 | 0.96 |
| +40 vs. +10 $< 0$ | -2.55 | 1.10 | [-4.38, -0.2] | 25 | 0.96 |
| +40 vs. baseline $< 0$ | -3.81 | 1.21 | [-5.8, -1.29] | 65 | 0.98 |
| **Exposure block 2** | | | | | |
| +10 vs. baseline $< 0$ | -1.41 | 0.69 | [-2.57, -0.21] | 29 | 0.97 |
| +40 vs. +10 $< 0$ | -3.46 | 0.88 | [-4.94, -1.64] | 100 | 0.99 |
| +40 vs. baseline $< 0$ | -4.86 | 1.00 | [-6.57, -2.75] | 257 | 1.00 |
| **Exposure block 3** | | | | | |
| +10 vs. baseline $< 0$ | -0.98 | 0.66 | [-2.15, 0.2] | 11 | 0.92 |
| +40 vs. +10 $< 0$ | -4.09 | 0.90 | [-5.61, -2.09] | 91 | 0.99 |
| +40 vs. baseline $< 0$ | -5.07 | 1.01 | [-6.81, -2.78] | 194 | 1.00 |

### §5.3.2 Simple effects of block within each exposure condition

Table S4
*Was there incremental change from exposure block 1 to 3? This table summarizes the simple effects of block for each exposure condition. Note that rightward shifts of the categorization function (and its PSE) correspond to negative estimates (lower intercepts in predicting the log-odds of "t"-responses).*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Difference between blocks: baseline** | | | | | |
| Block 1 to 2: decreased PSE | 0.15 | 0.53 | [-0.86, 1.17] | 1.6 | 0.61 |
| Block 2 to 3: decreased PSE | -0.04 | 0.47 | [-0.96, 0.8] | 0.9 | 0.47 |
| *Block 1 to 3: decreased PSE* | 0.11 | 0.68 | [-1.24, 1.39] | 1.3 | 0.57 |
| **Difference between blocks: +10** | | | | | |
| Block 1 to 2: decreased PSE | 0.11 | 0.67 | [-1.2, 1.38] | 1.3 | 0.57 |
| Block 2 to 3: decreased PSE | 0.42 | 0.61 | [-0.7, 1.53] | 3.0 | 0.75 |
| *Block 1 to 3: decreased PSE* | 0.53 | 0.84 | [-1.17, 2.14] | 2.7 | 0.73 |
| **Difference between blocks: +40** | | | | | |
| Block 1 to 2: increased PSE | -0.36 | 0.83 | [-1.82, 1.26] | 2.0 | 0.67 |
| Block 2 to 3: increased PSE | -0.12 | 0.73 | [-1.42, 1.18] | 1.3 | 0.57 |
| *Block 1 to 3: increased PSE* | -0.49 | 1.06 | [-2.34, 1.47] | 2.1 | 0.67 |

### §5.3.3 Differences in the rate of change between exposure conditions and block

Table S5
*Did the rate of block-to-block changes in PSEs differ between exposure conditions? This table summarizes the interactions between exposure condition and block—specifically, whether the differences between exposure conditions changed from exposure block to exposure block.*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Difference in +10 vs. baseline** | | | | | |
| Block 1 to 2: increased $\Delta_{PSE}$ | 0.06 | 0.61 | [-1.02, 1.16] | 0.9 | 0.46 |
| Block 2 to 3: increased $\Delta_{PSE}$ | 0.29 | 0.59 | [-0.76, 1.34] | 0.5 | 0.31 |
| *Block 1 to 3: increased* $\Delta_{PSE}$ | 0.34 | 0.79 | [-1.05, 1.76] | 0.5 | 0.33 |
| **Difference in +40 vs. +10** | | | | | |
| Block 1 to 2: increased $\Delta_{PSE}$ | -0.47 | 1.07 | [-2.35, 1.48] | 2.1 | 0.67 |
| Block 2 to 3: increased $\Delta_{PSE}$ | -0.53 | 0.82 | [-2, 0.97] | 2.9 | 0.74 |
| *Block 1 to 3: increased* $\Delta_{PSE}$ | -1.01 | 1.25 | [-3.31, 1.39] | 3.6 | 0.78 |
| **Difference in +40 vs. baseline** | | | | | |
| Block 1 to 2: increased $\Delta_{PSE}$ | -0.41 | 1.14 | [-2.43, 1.71] | 1.8 | 0.64 |
| Block 2 to 3: increased $\Delta_{PSE}$ | -0.47 | 1.11 | [-2.42, 1.49] | 2.0 | 0.66 |
| *Block 1 to 3: increased* $\Delta_{PSE}$ | -0.65 | 1.43 | [-3.25, 2.09] | 2.1 | 0.68 |

## §5.4 Slope results for test blocks

In the main text, we focused on 'shifts' in listeners' categorization function—i.e., changes in listeners' PSEs. Changes in the *slope* of the categorization function, or lack thereof, have received comparatively little attention in previous work (though see Clayards et al., 2008; Theodore & Monto, 2019). They are, however, an important part of the empirical facts that theories of speech perception need to account for (see also Kleinschmidt, 2020). While the present experiment was not primarily intended to test hypotheses about slope changes, we make a couple of post-hoc observations based on Bayesian hypothesis tests that parallel those for PSEs. Tables S6-S8 summarize Bayesian hypothesis tests about slopes that compare differences between conditions within blocks (Table S6), changes between blocks within conditions (Table S7), and interactions of exposure conditions and blocks (Table S8).

Our exposure conditions only manipulated the *absolute location* of category means, while holding constant both the relative distance between category means and the relative distance of each exposure token from those means (and thus also the category variances). Any model of adaptive speech perception that relies on distributional learning would thus predict little to no effect of exposure condition on the slope of listeners' categorization functions.[S7] This includes, for example, exemplar models (Johnson, 1997) and Bayesian ideal adaptors (Kleinschmidt & Jaeger, 2015).

This prediction was confirmed both prior to exposure and following exposure. Paralleling PSEs, there was evidence that slopes did not differ prior to exposure (top of Table S6), though the strength of this evidence was at best anecdotal, and thus weaker than for PSEs. The weak support for the null is not particularly surprising given the fact that we used regularizing priors. Such priors have their highest density over the null, weakening the ability to detect support for the null through the Savage-Dickey method (which compare the in/decrease of the posterior, compared to the prior, density over the null). Additionally, there are at least two *a priori* reasons

[S7] If our experiment only contained exposure blocks, no effect of exposure condition would be predicted. However, our experiment also contained test blocks, which did *not* differ between exposure conditions. To the extent that listeners attributed test trials to category /d/ or /t/, this can introduce differences in the category variances across exposure conditions. Overall, we would expect these differences to be small for two reasons: (1) test trials were unlabeled, leaving listeners with uncertainty as to *which* category to attribute them to, and (2) test blocks were shorter than exposure blocks (by the start of Test 4, listeners had heard 36 test trials and 144 exposure trials).

to expect that our paradigm would lead to high estimation uncertainty for the slopes (estimation of which the paradigm was not intended to prioritize). First, the VOT steps during test were at least 5ms apart. For subjects that have highly categorical perception—as was the case for many of our subjects—this makes it difficult to precisely estimate the actual slope. Second, our analysis approach, which assumed *linear* effects of VOT. Contrary to this simplifying assumption, there are reasons to expect quadratic effects of VOT (e.g., Bicknell et al., under review; Bushong & Jaeger, 2019, under review). It's possible that our simplifying assumption introduced additional estimation uncertainty.

In contrast to PSEs, there was also little evidence that slopes differed between exposure conditions at any point after exposure (remainder of Table S6). Compared to the changes in PSEs (Figure 5C), the *slopes* of listeners' categorization functions were similar across exposure conditions: the 95% CIs of the differences across conditions included 0 for all comparisons and all test blocks. BFs favored the null in all but two comparisons, though support for the null was again never more than anecdotal.

#### §5.4.1 Simple effects of condition within each test block

With regard to how categorization slopes (across all conditions) would change relative to the *pre*-exposure test, our expectations were less clear until we constructed the idealized learner reference lines in Figure 5C (after the experiment). On the one hand, we had designed the category variances of /d/ and /t/ to be somewhat plausible given the distribution VOT in American English. We did, for example, make sure that /t/ had larger VOT variance than /d/. So, to the extent that we created category variances that successfully resembled those that participants would expected based on their prior experience, exposure should not result in changes in participants' categorization slopes. On the other hand, methodological limitations (see Procedure in main text) kept us from most closely mimicking the distribution of VOTs in the database of American English that guided our design (Chodroff & Wilson, 2018, as described in §4.1).

#### §5.4.2 Simple effects of block within each exposure condition

Table S6

*Did exposure condition affect the slope of categorization responses? This table summarizes the differences in slopes between the exposure conditions within each test block. Since there was little reason to expect differences across conditions, all hypothesis tests contained in this table test the null (using the Savage-Dickey density ratio). We note that such tests depend on the prior.*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Test block 1 (pre-exposure)** | | | | | |
| +10 vs. baseline = 0 | -0.60 | 1.7 | [-4.39, 2.99] | 1.4 | 0.59 |
| +40 vs. +10 = 0 | 0.06 | 1.5 | [-3.18, 3.33] | 1.8 | 0.64 |
| +40 vs. baseline = 0 | -0.50 | 2.1 | [-5, 3.74] | 2.0 | 0.67 |
| **Test block 2** | | | | | |
| +10 vs. baseline = 0 | 1.28 | 1.9 | [-2.53, 5.64] | 1.1 | 0.54 |
| +40 vs. +10 = 0 | -2.30 | 2.2 | [-7.27, 1.8] | 0.7 | 0.42 |
| +40 vs. baseline = 0 | -1.02 | 2.5 | [-6.48, 4.13] | 1.5 | 0.60 |
| **Test block 3** | | | | | |
| +10 vs. baseline = 0 | 0.37 | 1.5 | [-2.73, 3.68] | 1.7 | 0.64 |
| +40 vs. +10 = 0 | -0.25 | 1.5 | [-3.4, 3.01] | 1.8 | 0.64 |
| +40 vs. baseline = 0 | 0.16 | 2.0 | [-3.78, 4.36] | 2.1 | 0.68 |
| **Test block 4** | | | | | |
| +10 vs. baseline = 0 | 1.26 | 1.8 | [-2.18, 5.09] | 1.2 | 0.55 |
| +40 vs. +10 = 0 | -0.02 | 1.7 | [-3.5, 3.68] | 1.6 | 0.62 |
| +40 vs. baseline = 0 | 1.33 | 2.1 | [-3.15, 5.91] | 1.5 | 0.61 |
| **Test block 5 (repeated testing without additional exposure)** | | | | | |
| +10 vs. baseline = 0 | 0.63 | 1.7 | [-3.06, 4.36] | 1.6 | 0.61 |
| +40 vs. +10 = 0 | 0.15 | 1.7 | [-3.45, 3.89] | 1.6 | 0.62 |
| +40 vs. baseline = 0 | 1.46 | 2.5 | [-3.69, 6.68] | 1.6 | 0.61 |
| **Test block 6 (repeated testing without additional exposure)** | | | | | |
| +10 vs. baseline = 0 | -1.50 | 2.0 | [-6.09, 2.38] | 1.1 | 0.52 |
| +40 vs. +10 = 0 | -1.68 | 1.8 | [-5.37, 1.74] | 1.0 | 0.50 |
| +40 vs. baseline = 0 | -0.40 | 2.5 | [-5.51, 4.7] | 1.8 | 0.64 |

### §5.4.3 Difference in the rate of change between exposure conditions and test block

Table S7
*Did participants' categorization slopes change between test blocks? This table summarizes the interactions between exposure condition and block—specifically whether the differences in slopes between exposure condition changed between blocks.*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Difference between blocks: baseline** | | | | | |
| Block 1 to 2: increased slope | 0.29 | 1.8 | [-3.07, 3.59] | 1.3 | 0.56 |
| Block 2 to 3: increased slope | -1.25 | 1.8 | [-4.46, 1.77] | 0.3 | 0.24 |
| Block 3 to 4: increased slope | 0.17 | 1.5 | [-2.39, 2.67] | 1.2 | 0.54 |
| *Block 1 to 4: increased slope* | -0.82 | 2.7 | [-5.37, 3.63] | 0.6 | 0.38 |
| Block 4 to 5: increased slope | 0.87 | 1.6 | [-1.76, 3.78] | 2.5 | 0.71 |
| Block 5 to 6: increased slope | 2.99 | 2.2 | [-0.43, 7.01] | 11.9 | 0.92 |
| *Block 4 to 6: increased slope* | 3.92 | 2.5 | [-0.1, 8.33] | 17.5 | 0.95 |
| **Difference between blocks: +10** | | | | | |
| Block 1 to 2: increased slope | 2.93 | 2.1 | [-0.41, 6.82] | 12.8 | 0.93 |
| Block 2 to 3: increased slope | -1.94 | 1.9 | [-5.69, 1.18] | 0.2 | 0.15 |
| Block 3 to 4: increased slope | 0.71 | 1.7 | [-2.04, 3.76] | 2.0 | 0.67 |
| *Block 1 to 4: increased slope* | 1.72 | 2.7 | [-2.82, 6.17] | 2.8 | 0.74 |
| Block 4 to 5: increased slope | -0.35 | 1.6 | [-3.27, 2.44] | 0.7 | 0.42 |
| Block 5 to 6: increased slope | 0.47 | 1.7 | [-2.41, 3.39] | 1.6 | 0.61 |
| *Block 4 to 6: increased slope* | 0.09 | 2.2 | [-3.56, 3.83] | 1.1 | 0.52 |
| **Difference between blocks: +40** | | | | | |
| Block 1 to 2: increased slope | 0.50 | 1.6 | [-2.13, 3.26] | 1.7 | 0.63 |
| Block 2 to 3: increased slope | 0.39 | 1.6 | [-2.28, 3.11] | 1.5 | 0.60 |
| Block 3 to 4: increased slope | 1.04 | 1.6 | [-1.67, 3.87] | 2.9 | 0.75 |
| *Block 1 to 4: increased slope* | 2.00 | 2.3 | [-1.88, 5.85] | 4.0 | 0.80 |
| Block 4 to 5: increased slope | -0.39 | 1.6 | [-3.2, 2.36] | 0.7 | 0.41 |
| Block 5 to 6: increased slope | -0.92 | 1.7 | [-3.87, 1.83] | 0.4 | 0.28 |
| *Block 4 to 6: increased slope* | -1.39 | 2.1 | [-4.88, 2.16] | 0.3 | 0.26 |

## §5.5 Slope results for exposure blocks

### §5.5.1 Simple effects of condition within each exposure block

### §5.5.2 Simple effects of block within each exposure condition

Table S8

*Did the slope differences between exposure conditions change from block to block? This table summarizes the interactions between exposure condition, block, and VOT—specifically, whether the differences in slopes between exposure conditions changed from test block to test block.*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Difference in slopes: +10 vs. baseline** | | | | | |
| Block 1 to 2: increased $\Delta_{slope}$ | 1.35 | 2.0 | [-1.95, 5.04] | 3.2 | 0.76 |
| Block 2 to 3: increased $\Delta_{slope}$ | -0.34 | 1.7 | [-3.34, 2.65] | 0.7 | 0.42 |
| Block 3 to 4: increased $\Delta_{slope}$ | 0.18 | 1.6 | [-2.57, 3.03] | 1.2 | 0.54 |
| *Block 1 to 4: increased* $\Delta_{slope}$ | 1.24 | 2.7 | [-3.35, 5.81] | 2.1 | 0.68 |
| Block 4 to 5: decreased $\Delta_{slope}$ | -0.67 | 1.7 | [-3.62, 2.16] | 1.9 | 0.65 |
| Block 5 to 6: decreased $\Delta_{slope}$ | -2.25 | 2.2 | [-6.15, 1.11] | 6.3 | 0.86 |
| *Block 4 to 6: decreased* $\Delta_{slope}$ | -2.98 | 2.6 | [-7.25, 1.11] | 7.7 | 0.88 |
| **Difference in slopes: +40 vs. +10** | | | | | |
| Block 1 to 2: increased $\Delta_{slope}$ | -1.50 | 2.0 | [-5.03, 1.54] | 0.3 | 0.21 |
| Block 2 to 3: increased $\Delta_{slope}$ | 1.67 | 1.9 | [-1.27, 5.06] | 4.7 | 0.82 |
| Block 3 to 4: increased $\Delta_{slope}$ | 0.63 | 1.7 | [-2.19, 3.72] | 1.8 | 0.64 |
| *Block 1 to 4: increased* $\Delta_{slope}$ | 0.84 | 2.7 | [-3.49, 5.22] | 1.7 | 0.62 |
| Block 4 to 5: decreased $\Delta_{slope}$ | -0.25 | 1.6 | [-3.24, 2.51] | 1.3 | 0.56 |
| Block 5 to 6: decreased $\Delta_{slope}$ | -1.72 | 1.9 | [-5.23, 1.29] | 4.8 | 0.83 |
| *Block 4 to 6: decreased* $\Delta_{slope}$ | -2.12 | 2.3 | [-6.04, 1.79] | 4.6 | 0.82 |
| **Difference in slopes: +40 vs. baseline** | | | | | |
| Block 1 to 2: increased $\Delta_{slope}$ | -0.17 | 2.5 | [-4.55, 4.16] | 0.9 | 0.47 |
| Block 2 to 3: increased $\Delta_{slope}$ | 1.41 | 2.4 | [-2.51, 5.36] | 2.7 | 0.73 |
| Block 3 to 4: increased $\Delta_{slope}$ | 0.86 | 2.1 | [-2.8, 4.46] | 1.9 | 0.66 |
| *Block 1 to 4: increased* $\Delta_{slope}$ | 2.18 | 3.4 | [-3.76, 7.83] | 2.8 | 0.73 |
| Block 4 to 5: decreased $\Delta_{slope}$ | -0.95 | 2.1 | [-4.57, 2.66] | 0.5 | 0.33 |
| Block 5 to 6: decreased $\Delta_{slope}$ | -4.18 | 2.6 | [-8.5, 0.12] | 0.1 | 0.06 |
| *Block 4 to 6: decreased* $\Delta_{slope}$ | -5.13 | 3.1 | [-10.32, 0.05] | 0.1 | 0.05 |

### §5.5.3 Differences in the rate of change between exposure conditions and block

Table S9
*Did exposure condition affect the slope of categorization responses? This table summarizes the differences in slopes between the exposure conditions within each exposure block*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Exposure block 1** | | | | | |
| +10 vs. baseline < 0 | 0.61 | 1.7 | [-2.22, 3.4] | 1.8 | 0.65 |
| +40 vs. +10 < 0 | 2.11 | 2.0 | [-1.02, 5.84] | 6.4 | 0.86 |
| +40 vs. baseline < 0 | 2.80 | 2.4 | [-1.18, 6.96] | 7.3 | 0.88 |
| **Exposure block 2** | | | | | |
| +10 vs. baseline < 0 | 0.69 | 1.8 | [-2.27, 3.88] | 1.9 | 0.66 |
| +40 vs. +10 < 0 | -0.89 | 1.9 | [-4.17, 2.26] | 0.5 | 0.32 |
| +40 vs. baseline < 0 | -0.13 | 2.4 | [-4.24, 3.82] | 0.9 | 0.48 |
| **Exposure block 3** | | | | | |
| +10 vs. baseline < 0 | -1.29 | 2.0 | [-4.9, 1.92] | 0.3 | 0.25 |
| +40 vs. +10 < 0 | 0.31 | 1.9 | [-2.99, 3.85] | 1.3 | 0.57 |
| +40 vs. baseline < 0 | -5.33 | 2.3 | [-9.4, -1.44] | 0.0 | 0.02 |

Table S10
*Did participants' categorization slopes change between exposure blocks? This table summarizes the interactions between exposure condition and block — specifically whether the differences in slopes between exposure condition changed between exposure blocks*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Difference between blocks: baseline** | | | | | |
| Block 1 to 2: increased slope | 1.43 | 1.7 | [-1.38, 4.49] | 4.1 | 0.80 |
| Block 2 to 3: increased slope | 1.64 | 1.9 | [-1.4, 5.1] | 4.4 | 0.81 |
| Block 1 to 3: increased slope | 3.17 | 2.5 | [-1.07, 7.45] | 8.5 | 0.89 |
| **Difference between blocks: +10** | | | | | |
| Block 1 to 2: increased slope | 1.72 | 1.8 | [-1.18, 4.98] | 4.9 | 0.83 |
| Block 2 to 3: increased slope | -0.81 | 1.7 | [-3.85, 2.11] | 0.5 | 0.33 |
| Block 1 to 3: increased slope | 0.96 | 2.3 | [-2.86, 4.82] | 1.9 | 0.65 |
| **Difference between blocks: +40** | | | | | |
| Block 1 to 2: increased slope | -1.53 | 2.0 | [-5.31, 1.72] | 0.3 | 0.21 |
| Block 2 to 3: increased slope | -0.17 | 1.8 | [-3.27, 2.99] | 0.9 | 0.46 |
| Block 1 to 3: increased slope | -1.70 | 2.6 | [-6.36, 2.53] | 0.3 | 0.25 |



## §5.6 Lapse rates by exposure and test blocks

All analyses presented in the main text and in the preceding SI sections assume a constant lapse rate across exposure and test blocks. These analyses found lapse rates of 0.8% (95%-CI: NA%-NA%) during test blocks and 0.1% (95%-CI: 0%-0.4%) during exposure blocks, much smaller than in previous work (Clayards et al., 2008; Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016). The

Table S11

*Did the rate of block-to-block changes differ across exposure conditions? This table summarizes the interactions between between VOT, exposure condition, and block —specifically, whether the differences between exposure conditions changed from exposure block to exposure block.*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Difference in +10 vs. baseline** | | | | | |
| Block 1 to 2: increased $\Delta_{slope}$ | -0.33 | 1.7 | [-3.26, 2.6] | 1.4 | 0.58 |
| Block 2 to 3: increased $\Delta_{slope}$ | -1.51 | 1.9 | [-4.98, 1.59] | 3.7 | 0.79 |
| *Block 1 to 3: increased* $\Delta_{slope}$ | -1.88 | 2.5 | [-6.04, 2.12] | 3.6 | 0.78 |
| **Difference in +40 vs. +10** | | | | | |
| Block 1 to 2: increased $\Delta_{slope}$ | -2.78 | 2.5 | [-7.31, 0.82] | 8.2 | 0.89 |
| Block 2 to 3: increased $\Delta_{slope}$ | 0.23 | 1.9 | [-3.01, 3.69] | 0.8 | 0.45 |
| *Block 1 to 3: increased* $\Delta_{slope}$ | -2.59 | 2.9 | [-7.73, 2.17] | 4.5 | 0.82 |
| **Difference in +40 vs. baseline** | | | | | |
| Block 1 to 2: increased $\Delta_{slope}$ | -3.28 | 2.6 | [-7.86, 1.03] | 8.3 | 0.89 |
| Block 2 to 3: increased $\Delta_{slope}$ | -0.10 | 2.7 | [-4.5, 4.57] | 1.1 | 0.52 |
| *Block 1 to 3: increased* $\Delta_{slope}$ | -4.62 | 3.4 | [-10.28, 1.22] | 9.7 | 0.91 |

decision to assume constant lapse rates seems to be supported by Figures 5A-B, which show that participants' responses (the point ranges) in all exposure and test blocks approach 0 and 100% "t"-responses for small and large VOTs, respectively. Still, to explore changes in lapse rates we refitted the same psychometric mixed-effects model with a lapse model that included block as a predictor (sliding-difference coded from the first to the ninth block). Priors in this model were the same as that in the constant-lapse model including the prior for lapse rates which was the `brms` (Bürkner, 2017) default of a logistic prior centered around 0 with a scale of 1 unit. This prior assigns fairly equal weighting over values of $0 < p < 1$ and lower weighting of the extreme values.

The estimated lapse rate for each block is shown in Table S12. Lapse rates were at their highest in the first block (1.5%) but fell sharply up to the fifth block ($\leq$ .0011%) before rising in the sixth (0.07%) and seventh block (0.04%). Where estimates were especially low, we note the accompanying wide confidence intervals partly resulting from our use of a uniform prior. Overall, these results show some evidence of familiarization with the task. However, unlike previous work, lapse rates were very low even during the first 12 trials (the first test block). For comparison, a post-hoc analysis reported in Kleinschmidt (2020) estimated lapse rates as high as 12% during the first 37 trials of his experiment. Lapse rates reduced to about 5% over the remaining 185 trials.

Table S12
*This table summarizes the estimated proportion of lapses for each block of the experiment.*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| Block 1 : $\lambda > 0$ | 0.02 | 0.01 | [0.005, 0.04] | $\geq 8000$ | 1 |
| Block 2 : $\lambda > 0$ | 0.00 | 0.00 | [0, 0.003] | $\geq 8000$ | 1 |
| Block 3 : $\lambda > 0$ | 0.00 | 0.00 | [0, 0.007] | $\geq 8000$ | 1 |
| Block 4 : $\lambda > 0$ | 0.00 | 0.00 | [0, 0.003] | $\geq 8000$ | 1 |
| Block 5 : $\lambda > 0$ | 0.00 | 0.00 | [0, 0.012] | $\geq 8000$ | 1 |
| Block 6 : $\lambda > 0$ | 0.00 | 0.00 | [0, 0.005] | $\geq 8000$ | 1 |
| Block 7 : $\lambda > 0$ | 0.01 | 0.00 | [0.001, 0.017] | $\geq 8000$ | 1 |
| Block 8 : $\lambda > 0$ | 0.00 | 0.00 | [0, 0.004] | $\geq 8000$ | 1 |
| Block 9 : $\lambda > 0$ | 0.01 | 0.01 | [0, 0.022] | $\geq 8000$ | 1 |

One possible explanation for the consistently small lapse rate in the present experiment is that we recruited participants from the Prolific crowdsourcing platform rather than Mechanical Turk (Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016). Prolific has been found to deliver higher data quality than the Amazon's Mechanical Turk crowd-sourcing platform (Adams, Li, & Liu, 2020; Albert & Smilek, 2023; Peer, Brandimarte, Samat, & Acquisti, 2017). Another possibility is that the use of natural-, rather than robotic-sounding, stimuli engaged participants attention in the present experiment.

## §5.7 Changes in participants' recognition accuracy as a function of exposure

Figure S6 summarizes changes in participants' recognition accuracy with increasing exposure to the different exposure conditions. The figure reproduces Figure 7 while additionally showing listeners' empirical recognition accuracy, as estimated from the unlabeled trials of exposure blocks.

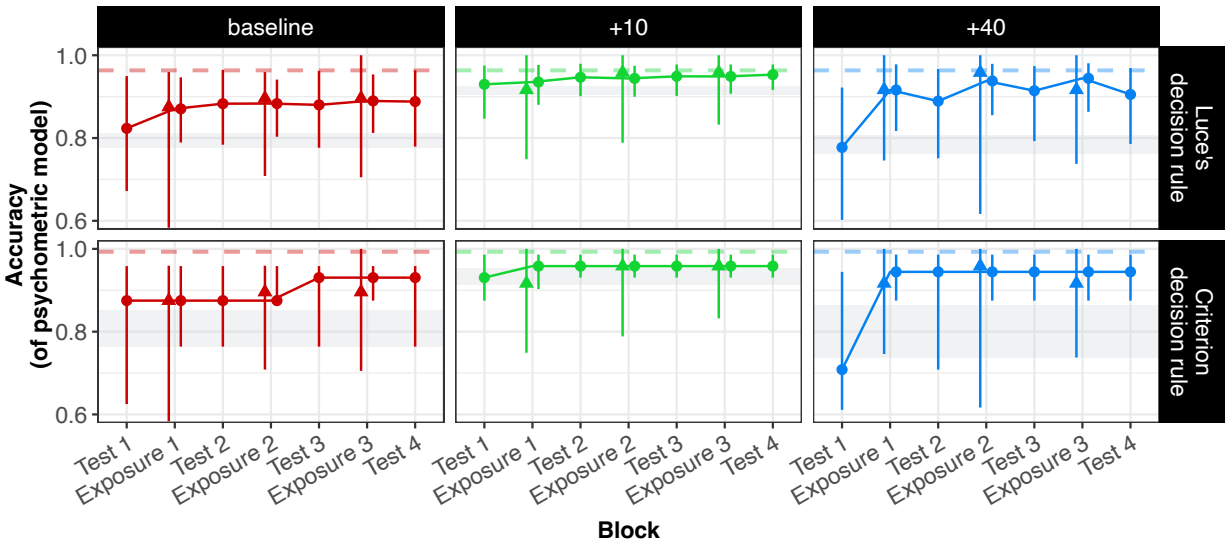


*Figure S6*. Same as Figure 7 but also showing listeners' empirical recognition accuracy (triangles). For the empirical accuracy, lineranges show 95% bootstrapped CIs.

## §5.8 Comparing changes in participants' behavior against idealized learner models

This section elaborates on Table 4 in the main text, comparing changes in participants' behavior against idealized learner models. In the main text, we reported that participants' behavior in Test 4 had not converged against the behavior of an idealized learner, who has fully learned the exposure distributions. Here, we first report the same test for all exposure and test blocks, confirming the lack of full learning for all blocks.

Then we report hypothesis tests that assess whether the *degree* of convergence—quantified by the percentage labels in Figure 5C in the main text— differed between exposure conditions. We are aware of only two previous studies that speak to this question (Kleinschmidt, 2020; Kleinschmidt & Jaeger, 2016). Both studies compared the degree of convergence between multiple

exposure distributions that were left- or rightward-shifted along the VOT continuum. Both studies observed the same two trends we observe here: (1) the degree of convergence was proportionally lower for more extreme shifts (see baseline vs. +10 condition in Test 4) and (2) the degree of convergence was lower for shifts that were leftward rather than rightward (baseline vs. +40 condition, which are shifted about equally far relative to listeners' prior expectations). We refer to finding (1) as *shrinkage* and to finding (2) as *asymmetric shrinkage*. However, neither of the two previous studies assessed the credibility of findings (1) and (2). As we detail below, such differences—if confirmed—would have consequences for theories of adaptive speech perception.

#### §5.8.1 Testing convergence against idealized learner model

Table S13 reproduces Table 4 for all exposure and test blocks.

### §5.8.2 Testing 'shrinkage': Do exposure conditions differ in the degree to which they converged against idealized learner models?

Recall that all three exposure conditions had the expected qualitative effects both relative to each other, and relative to participants' pre-exposure expectations: the absolute magnitude of the shift qualitatively followed the predictions of distributional learning models. This leaves open, however, whether the *degree* of adaptation was comparable across exposure conditions. That is, after the same amount of exposure, did participants in exposure conditions with more extreme shifts (e.g., the baseline condition) exhibit *proportionally* smaller changes in their categorization functions, compared to participants in exposure conditions with less extreme shifts (e.g., the +0 condition)? To assess whether adaptive changes in participants' behavior exhibited such 'shrinkage', it is necessary to quantify the degree to which participants in different exposure conditions converged against the categorization functions expected from idealized learners. Here we do so by estimating the distance of participants' PSE from the PSE expected prior to exposure, and then comparing this distance to the distance of the ideal learner's PSE from that same prior PSE:

$$\text{Proportion convergence towards PSE of idealized learner} = \frac{\Delta(PSE_{actual}, PSE_{pre-exposure})}{\Delta(PSE_{ideal}, PSE_{pre-exposure})} \tag{S1}$$

This proportion can then be compared across different exposure conditions, allowing us to ask two questions of theoretical relevance.

1. **Is there shrinkage?** Is proportion convergence smaller for the more extreme leftward-shifted baseline condition, compared to the less extreme leftward-shifted +10 condition?
2. **Is there asymmetric shrinkage?** Is proportion convergence smaller for the *left*ward-shifted baseline condition, compared to the equally extreme *right*ward-shifted +40 condition?

However, equation ((S1)) contains a researcher degree of freedom: the approach to estimating $PSE_{pre-exposure}$. Here, we entertained three different ways of estimating

$PSE_{pre-exposure}$, all of which yield similar results: anecdotal, and by no means convincing, evidence of asymmetric shrinkage. The hypothesis tests for the three different approaches are summarized in Tables S14 to S16. Table S14 compares participants' actual PSE (and the idealized learner's PSE) against the pre-exposure PSE expected from an idealized pre-exposure listener. Table **??** compares against the PSE that participants in the same condition had prior to any exposure—i.e., participants' PSE in Test 1. **This is the approach we used to calculate the percentage labels in Figure 5C in the main text**, as it struck us to most straightforwardly account for potential individual differences between listeners. Finally, Table S16 compares against the participants' pre-exposure PSE in Test 1 while averaging over all three conditions.

### §5.8.3 Discussion of shrinkage

As mentioned above, we are aware of only two previous studies that measured the degree of convergence across multiple exposure conditions. Kleinschmidt and Jaeger (2016) exposed different groups of listeners to different VOT distributions for /b/ and /p/ that were shifted to different degrees left- or rightwards relative to listeners' prior expectations. An unpublished follow-up analysis of the same data (Kleinschmidt, 2020) revealed asymmetric shrinkage patterns that closely resemble those found here. First, larger shifts in the exposure distribution yielded shifts in listeners' categorization functions that were *proportionally smaller* relative to the shifts expected from an idealized learner (shrinkage). Second, shrinkage effects were more pronounced for leftward- than rightward-shifted exposure distribution (asymmetric shrinkage). Kleinschmidt (2020) also presents additional experiments with extreme examples of leftward-shifted exposures. In his Experiment 4, different groups of listeners were exposed to /b/-/p/ distributions for which the /b/ mean was shifted to -20ms, -50ms, or even -80ms, while the /p/ mean remained at 50ms VOT. This experiment replicated the strong shrinkage effects for such leftward shifts: even in the most extreme condition (-80ms), listeners' average PSE was still larger than 20ms.[S8] Across a total of ten left- and right-wards shifted exposure conditions, the findings in Kleinschmidt support both shrinkage and that this shrinkage is (much) more pronounced for left-wards shifts along the

[S8] This second set of results has to be interpreted with caution given potential issues with the data quality (cf. Kleinschmidt, 2020, p. 25).

VOT continuum. It is, however, important to emphasize that Kleinschmidt presents no statistical tests that assessed the credibility of this pattern.

As our own tests of the shrinkage hypothesis in Tables S14 to S16 show, even seemingly striking shrinkage patterns do not necessarily provide convincing evidence for shrinkage. In the present case, we find that the median degree of convergence at Test 4 relative to listeners' pre-exposure PSE ranged from 20.7% in the baseline condition to 34% in the +10 condition and 36.8% in the +40 condition. At first blush, this would seem to support rather large difference in the degree of convergence. For the two leftward-shifted exposure conditions, the more extreme baseline condition achieved less than 50% of the degree of convergence found in the less extreme +10 condition. And even though the baseline and +40 conditions were shifted about equally far relative to participants' prior expectations, the leftward-shifted baseline condition elicited in less than 50% of the degree of convergence observed for the rightward-shifted +40 condition. Together, this would seem to suggest shrinkage that is asymmetric, in that it is more pronounced for exposure conditions that leftward-shifted along the VOT continuum (relative to listeners' prior expectations). However, none of the three approaches considered above finds more than anecdotal evidence (BFs $< 3$) that this shrinkage pattern is credible.

We note this anecdotal trend here primarily because of its theoretical relevance, if found credible by future studies. Existing distributional learning models would *not* seem to predict **shrinkage** or **asymmetric shrinkage**. However, the model selection approach we discussed in the general discussion would arguably predict asymmetric shrinkage, provided sufficient statistical power to detect it. Specifically, most L1-US English listeners should be unlikely to shift their categorization functions particularly far *leftwards* because they are unlikely to have experienced many talkers with such leftward-shifted PSEs. This asymmetry is due to the fact that the distribution of VOT values is bounded at zero (see Figure 6), leaving only very limited room for leftward shifts along the VOT continuum.[S9] In other words, how talkers of US English realize

[S9] Readers familiar with the literature on syllable-initial stop-voicing might object that pre-voicing—sometimes treated as negative VOT—even occurs in languages like English, in which it is not considered a primary cue to stop-voicing (Chodroff & Wilson, 2018 did not annotate pre-voicing). However, it is an empirical question whether listeners treat pre-voicing as a separate phonetic cue, in which case VOT would be bounded at zero (by definition). Even if listeners interpret pre-voicing as negative VOT, this does not change that L1-US English talkers' categorization functions have very little room to move leftwards along the VOT continuum. This is the case because even L1-US English talkers that frequently produce pre-voicing tend to *also* produce positive VOTs for the voiced

syllable-initial stop-voicing contrasts imposes a strong constraint on the extent to which listeners can shift their categorization functions leftwards along the VOT continuum—at least if rapid adaptation during the initial moments of encountering an unfamiliar talker relies primarily on model selection. This predicted inability to accommodate leftward shifts along the VOT continuum should be particularly pronounced for the bilabial stops (/b/-/p/), as the VOT distributions of /b/ is particular close to 0.

The model selection approach to distributional learning also predicts that listeners who have ample experience with accents that more systematically prevoice syllable-initial voiced stops (e.g., L2-accented English spoken by speakers of L1-Mexican Spanish) should show an increased flexibility to accommodate leftward shifts along the VOT continuum, and thus less asymmetric shrinkage. It is, however, important to keep in mind that the category boundaries for even these accents tend to still have positive VOTs, so that the predicted reduction in asymmetric shrinkage might be relative subtle and difficult to detect.

## §5.9 Relaxing the linearity assumption for VOT

The perceptual model of our psychometric mixed-effects analyses assumed linear effects of VOT on the log-odds of "t"-responses. As already mentioned in SI §4.3, we made this assumptions for the sake of simplicity, and in order to avoid over-fitting. However, there are reasons to expect that the effects of VOT on listeners' categorizations are non-linear. Specifically, ideal observers with Gaussian categories along a single cue dimension (here: VOT) predict that the posterior log-odds of each category change linearly along that cue dimension if and only if the variance of both Gaussian categories is equal (for details, see Bicknell et al., under review; Kleinschmidt & Jaeger, 2015; Kronrod et al., 2016). For US English /d/ and /t/, this is well-known *not* to be the case, with /t/ having larger variance than /d/ (see also Figure 6 in the main text). Under the assumption of Gaussian categories, ideal observer thus predict a quadratic effect along VOT—specifically, the log-odds of "t"-responses are predicted to increase more than linearly with

category (including /d/). That is, talkers like the synthetic talkers that Kleinschmidt (2020) created, who produce the voiced category with large negative VOTs (pre-voicing) but produce the voiceless category with large positive VOTs are rarely, if ever, experienced by a typical L1 listeners of US English. As a consequence, the category boundaries of the vast majority of talkers (and thus the ideal PSEs for speech from those talkers) that most L1 listeners of US English have experienced are positive.

increasing VOT between the two category means (since the /t/ category has larger variance).

While the linearity assumption made in our main analysis does not *necessarily* introduce statistical bias, it is possible that it does. We thus conducted additional analyses that relaxed the linearity assumption by modeling the effect of VOT as a non-parametric smooth. Since this analysis affords additional degrees of freedom, we reduced concerns about over-fitting by combining the (unlabeled) exposure and test data into a single analysis. Instead of modeling effects block by block, we decided to model the incremental and cumulative effects of exposure by including a non-linear effect of trial and its interaction with VOT in the analysis. In addition to replicating our analysis under relaxed assumptions about linearity, this auxiliary analysis also sheds light on the causes for the 'zigzag' pattern in the intercept and slope estimates for exposure and test blocks that we reported in the main analysis.

#### §5.9.1 Substituting GAMMs for GLMMs in our psychometric mixed-effects model

Specifically, we use the same psychometric mixed-effects model as in the main analysis, except that we replaced the full factorial of VOT, block, and exposure condition in the perceptual model with a separate tensor smooth of VOT and trial for each of the three exposure conditions.[S10]. This makes this auxiliary analysis a mixed-effect mixture model, for which the mixture component that is the perceptual model is a generalized additive mixed-effect models (GAMM), rather than a generalized linear mixed-effects model. The analysis contained the same full random effect structure and priors as the psychometric model presented in the main text. Both VOT and trial were Gelman-scaled prior to the analysis.

#### §5.9.2 Results

Figure S7 shows the predicted log-odds of "t"-responses that result from the fitted GAMM. The GAMM results replicate the directional effects of exposure both within and across exposure conditions. For the baseline condition, the contour lines largely shift downwards with exposure. This means that the same VOT is more likely to be categorized as "t" with increasing exposure.

[S10] `t2(VOT, Trial, by = Condition.Exposure, bs = "tp")`

For the +40 condition, the opposite trend is observed (and much more clearly), indicating that the same VOT is *less* likely to be categorized as "t" with increasing exposure. The +10 condition falls between the baseline and +40 condition.

The GAMM results also confirm that the changes introduced by exposure are undone with repeated testing. This shows in the contour lines over trials 160-175, which trend towards reverting the changes introduced by the preceding exposure. The trend shows most clearly for the baseline and +40 condition (in opposite directions since the exposure effects are in opposite directions for these two conditions).

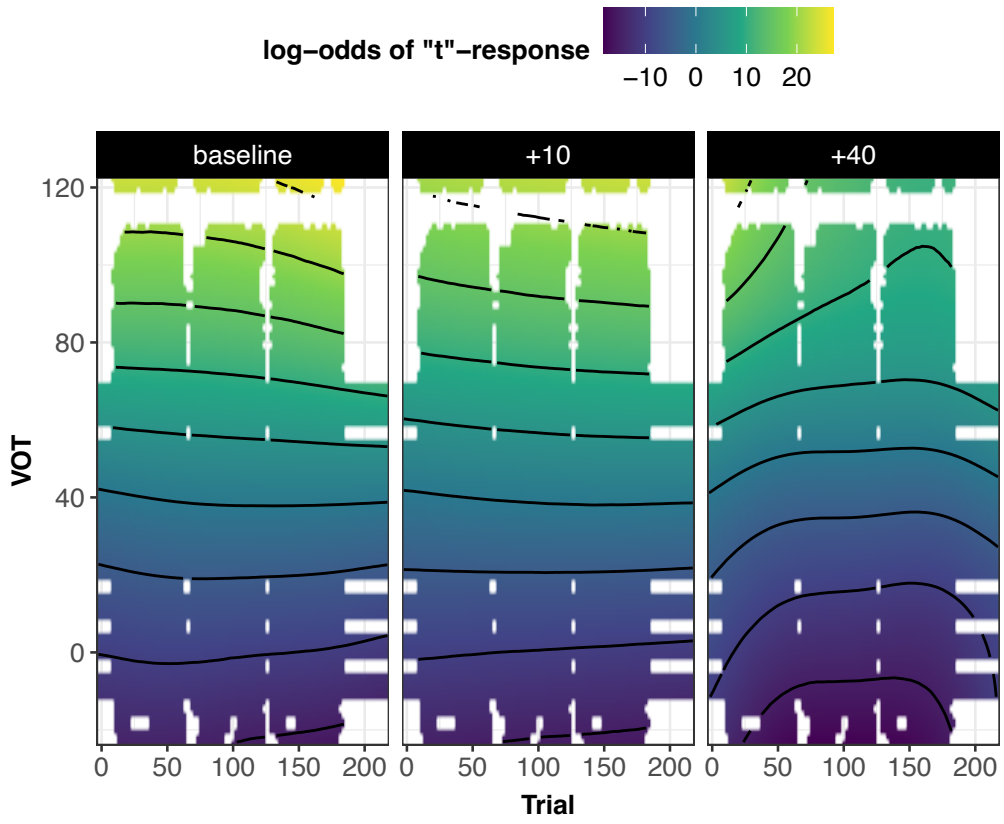


*Figure S7.* Predicted log-odds of "t"-response by trial (across all exposure and test blocks) and VOT for each of the three exposure conditions. Combinations of trial and VOT that did not occur in any of the three exposure conditions are left white (the reduced VOT range of test blocks makes them easily identifiable).

## §6 Visual analysis of reaction times

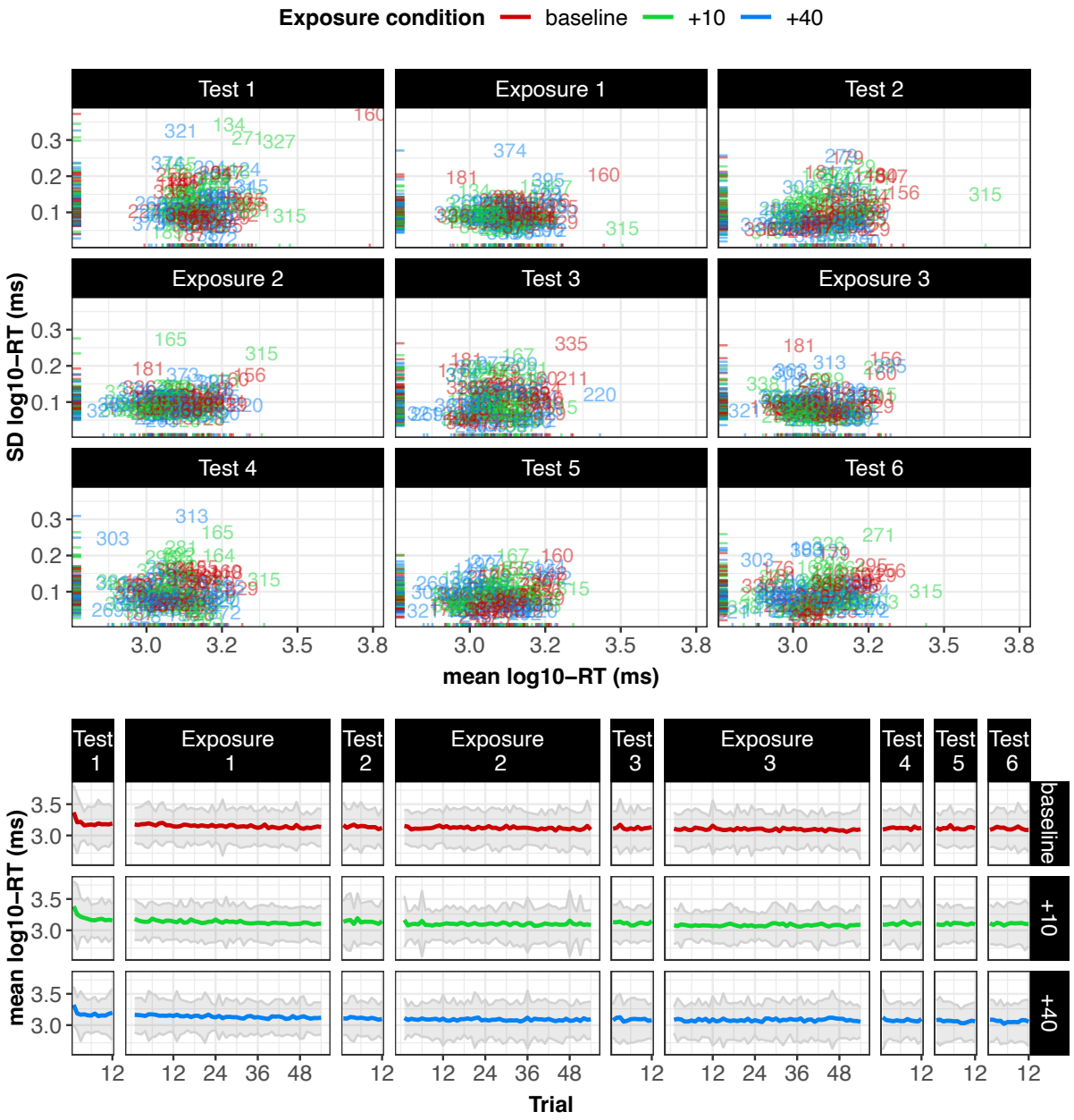


*Figure S8.* Summary of participant reaction times by block. **Top:** Scatter plot of means and SDs of RTs in $log_{10}$ms by participant and block. **Bottom:** Progression of RTs by trial (including catch trials) and block. Lines indicate mean RTs, ribbons indicate region of two times SD from the mean.

# §7 Comparing predictions of ideal adaptor against participants' responses

## §7.1 Fit with uninformative (regularizing) priors

Here some initial plots. While the plots are back-transformed into the original cue space, the table is not (yet).

## §7.2 Fit with informative priors about expected category means and covariances based on phonetic data from Chodroff and Wilson (2018)

# §8 Session Info

```
## - Session info ---------------------------------------------------------------------------
##  setting  value
##  version  R version 4.3.1 (2023-06-16)
##  os       macOS Monterey 12.7.2
##  system   aarch64, darwin20
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  ctype    en_US.UTF-8
##  tz       Europe/Stockholm
##  date     2024-05-08
##  pandoc   3.1.1 @ /Applications/RStudio.app/Contents/Resources/app/quarto/bin/tools/ (via rm
##
## - Packages -------------------------------------------------------------------------------
##  package         * version    date (UTC) lib source
##  abind             1.4-5      2016-07-21 [1] CRAN (R 4.3.0)
##  arrayhelpers      1.1-0      2020-02-04 [1] CRAN (R 4.3.0)
##  assertthat      * 0.2.1      2019-03-21 [1] CRAN (R 4.3.0)
```

```
##  av                0.9.0      2023-12-05 [1] CRAN (R 4.3.1)
##  backports         1.4.1      2021-12-13 [1] CRAN (R 4.3.0)
##  base64enc         0.1-3      2015-07-28 [1] CRAN (R 4.3.0)
##  bayesplot         1.11.1     2024-02-15 [1] CRAN (R 4.3.1)
##  bayestestR        0.13.2     2024-02-12 [1] CRAN (R 4.3.1)
##  bit               4.0.5      2022-11-15 [1] CRAN (R 4.3.0)
##  bit64             4.0.5      2020-08-30 [1] CRAN (R 4.3.0)
##  bookdown          0.39       2024-04-15 [1] CRAN (R 4.3.1)
##  boot              1.3-30     2024-02-26 [1] CRAN (R 4.3.1)
##  bridgesampling    1.1-2      2021-04-16 [1] CRAN (R 4.3.0)
##  brms            * 2.20.4     2023-09-25 [1] CRAN (R 4.3.1)
##  Brobdingnag       1.2-9      2022-10-19 [1] CRAN (R 4.3.0)
##  broom             1.0.5      2023-06-09 [1] CRAN (R 4.3.0)
##  broom.mixed     * 0.2.9.4    2022-04-17 [1] CRAN (R 4.3.0)
##  cachem            1.0.8      2023-05-01 [1] CRAN (R 4.3.0)
##  checkmate         2.3.1      2023-12-04 [1] CRAN (R 4.3.1)
##  class             7.3-22     2023-05-03 [1] CRAN (R 4.3.1)
##  classInt          0.4-10     2023-09-05 [1] CRAN (R 4.3.0)
##  cli               3.6.2      2023-12-11 [1] CRAN (R 4.3.1)
##  clue              0.3-65     2023-09-23 [1] CRAN (R 4.3.1)
##  cluster           2.1.6      2023-12-01 [1] CRAN (R 4.3.1)
##  cmdstanr          0.5.3      2023-06-26 [1] local
##  coda              0.19-4.1   2024-01-31 [1] CRAN (R 4.3.1)
##  codetools         0.2-19     2023-02-01 [1] CRAN (R 4.3.1)
##  colorspace        2.1-0      2023-01-23 [1] CRAN (R 4.3.0)
##  colourpicker      1.3.0      2023-08-21 [1] CRAN (R 4.3.0)
##  cowplot           1.1.3      2024-01-22 [1] CRAN (R 4.3.1)
##  crayon            1.5.2      2022-09-29 [1] CRAN (R 4.3.0)
##  crosstalk         1.2.1      2023-11-23 [1] CRAN (R 4.3.1)
##  curl              5.2.1      2024-03-01 [1] CRAN (R 4.3.1)
```

```
##  data.table        1.15.2    2024-02-29 [1] CRAN (R 4.3.1)
##  datawizard        0.9.1     2023-12-21 [1] CRAN (R 4.3.1)
##  DBI               1.2.2     2024-02-16 [1] CRAN (R 4.3.1)
##  devtools          2.4.5     2022-10-11 [1] CRAN (R 4.3.0)
##  digest            0.6.35    2024-03-11 [1] CRAN (R 4.3.1)
##  diptest         * 0.77-0    2023-11-27 [1] CRAN (R 4.3.1)
##  distributional    0.4.0     2024-02-07 [1] CRAN (R 4.3.1)
##  dplyr           * 1.1.4     2023-11-17 [1] CRAN (R 4.3.1)
##  DT                0.32      2024-02-19 [1] CRAN (R 4.3.1)
##  dygraphs          1.1.1.6   2018-07-11 [1] CRAN (R 4.3.0)
##  e1071             1.7-14    2023-12-06 [1] CRAN (R 4.3.1)
##  effectsize        0.8.6     2023-09-14 [1] CRAN (R 4.3.0)
##  ellipse           0.5.0     2023-07-20 [1] CRAN (R 4.3.0)
##  ellipsis          0.3.2     2021-04-29 [1] CRAN (R 4.3.0)
##  emmeans           1.10.0    2024-01-23 [1] CRAN (R 4.3.1)
##  estimability      1.5       2024-02-20 [1] CRAN (R 4.3.1)
##  evaluate          0.23      2023-11-01 [1] CRAN (R 4.3.1)
##  extraDistr        1.10.0    2023-11-30 [1] CRAN (R 4.3.1)
##  fansi             1.0.6     2023-12-08 [1] CRAN (R 4.3.1)
##  farver            2.1.1     2022-07-06 [1] CRAN (R 4.3.0)
##  fastmap           1.1.1     2023-02-24 [1] CRAN (R 4.3.0)
##  fBasics           4032.96   2023-11-03 [1] CRAN (R 4.3.1)
##  forcats         * 1.0.0     2023-01-29 [1] CRAN (R 4.3.0)
##  foreach           1.5.2     2022-02-02 [1] CRAN (R 4.3.0)
##  foreign           0.8-86    2023-11-28 [1] CRAN (R 4.3.1)
##  Formula           1.2-5     2023-02-24 [1] CRAN (R 4.3.0)
##  fs                1.6.4     2024-04-25 [1] CRAN (R 4.3.1)
##  furrr           * 0.3.1     2022-08-15 [1] CRAN (R 4.3.0)
##  future          * 1.33.1    2023-12-22 [1] CRAN (R 4.3.1)
##  generics          0.1.3     2022-07-05 [1] CRAN (R 4.3.0)
```

```
##  gganimate         1.0.9      2024-02-27 [1] CRAN (R 4.3.1)
##  ggdist            3.3.2      2024-03-05 [1] CRAN (R 4.3.1)
##  ggforce         * 0.4.2      2024-02-19 [1] CRAN (R 4.3.1)
##  ggnewscale      * 0.4.10     2024-02-08 [1] CRAN (R 4.3.1)
##  ggplot2         * 3.5.0      2024-02-23 [1] CRAN (R 4.3.1)
##  ggridges          0.5.6      2024-01-23 [1] CRAN (R 4.3.1)
##  ggstance        * 0.3.6      2022-11-16 [1] CRAN (R 4.3.0)
##  ggtext          * 0.1.2      2022-09-16 [1] CRAN (R 4.3.0)
##  gifski            1.12.0-2   2023-08-12 [1] CRAN (R 4.3.0)
##  globals           0.16.3     2024-03-08 [1] CRAN (R 4.3.1)
##  glue              1.7.0      2024-01-09 [1] CRAN (R 4.3.1)
##  gridExtra         2.3        2017-09-09 [1] CRAN (R 4.3.0)
##  gridtext          0.1.5      2022-09-16 [1] CRAN (R 4.3.0)
##  gtable            0.3.4      2023-08-21 [1] CRAN (R 4.3.0)
##  gtools            3.9.5      2023-11-20 [1] CRAN (R 4.3.1)
##  here            * 1.0.1      2020-12-13 [1] CRAN (R 4.3.0)
##  Hmisc             5.1-1      2023-09-12 [1] CRAN (R 4.3.0)
##  hms               1.1.3      2023-03-21 [1] CRAN (R 4.3.0)
##  htmlTable         2.4.2      2023-10-29 [1] CRAN (R 4.3.1)
##  htmltools         0.5.8.1    2024-04-04 [1] CRAN (R 4.3.1)
##  htmlwidgets       1.6.4      2023-12-06 [1] CRAN (R 4.3.1)
##  httpuv            1.6.14     2024-01-26 [1] CRAN (R 4.3.1)
##  igraph            2.0.2      2024-02-17 [1] CRAN (R 4.3.1)
##  inline            0.3.19     2021-05-31 [1] CRAN (R 4.3.0)
##  insight           0.19.8     2024-01-31 [1] CRAN (R 4.3.1)
##  isoband           0.2.7      2022-12-20 [1] CRAN (R 4.3.0)
##  iterators         1.0.14     2022-02-05 [1] CRAN (R 4.3.0)
##  jsonlite          1.8.8      2023-12-04 [1] CRAN (R 4.3.1)
##  kableExtra      * 1.4.0      2024-01-24 [1] CRAN (R 4.3.1)
##  KernSmooth        2.23-22    2023-07-10 [1] CRAN (R 4.3.0)
```

```
##  knitr             1.45       2023-10-30 [1] CRAN (R 4.3.1)
##  labeling          0.4.3      2023-08-29 [1] CRAN (R 4.3.0)
##  LaplacesDemon     16.1.6     2021-07-09 [1] CRAN (R 4.3.0)
##  later             1.3.2      2023-12-06 [1] CRAN (R 4.3.1)
##  latexdiffr      * 0.2.0      2024-02-16 [1] CRAN (R 4.3.1)
##  lattice           0.22-5     2023-10-24 [1] CRAN (R 4.3.1)
##  lifecycle         1.0.4      2023-11-07 [1] CRAN (R 4.3.1)
##  linguisticsdown * 1.2.0      2019-03-01 [1] CRAN (R 4.3.0)
##  listenv           0.9.1      2024-01-29 [1] CRAN (R 4.3.1)
##  lme4              1.1-35.1   2023-11-05 [1] CRAN (R 4.3.1)
##  loo               2.7.0      2024-02-24 [1] CRAN (R 4.3.1)
##  lpSolve           5.6.20     2023-12-10 [1] CRAN (R 4.3.1)
##  lubridate       * 1.9.3      2023-09-27 [1] CRAN (R 4.3.1)
##  magick          * 2.8.3      2024-02-18 [1] CRAN (R 4.3.1)
##  magrittr        * 2.0.3      2022-03-30 [1] CRAN (R 4.3.0)
##  markdown          1.12       2023-12-06 [1] CRAN (R 4.3.1)
##  MASS              7.3-60.0.1 2024-01-13 [1] CRAN (R 4.3.1)
##  Matrix            1.6-5      2024-01-11 [1] CRAN (R 4.3.1)
##  matrixStats       1.2.0      2023-12-11 [1] CRAN (R 4.3.1)
##  memoise           2.0.1      2021-11-26 [1] CRAN (R 4.3.0)
##  mgcv              1.9-1      2023-12-21 [1] CRAN (R 4.3.1)
##  mime              0.12       2021-09-28 [1] CRAN (R 4.3.0)
##  miniUI            0.1.1.1    2018-05-18 [1] CRAN (R 4.3.0)
##  minqa             1.2.6      2023-09-11 [1] CRAN (R 4.3.0)
##  modeest           2.4.0      2019-11-18 [1] CRAN (R 4.3.0)
##  multcomp          1.4-25     2023-06-20 [1] CRAN (R 4.3.0)
##  munsell           0.5.0      2018-06-12 [1] CRAN (R 4.3.0)
##  MVBeliefUpdatr  * 0.0.1.0005 2024-05-05 [1] Github (hlplab/MVBeliefUpdatr@bbfac50)
##  mvtnorm           1.2-4      2023-11-27 [1] CRAN (R 4.3.1)
##  nlme              3.1-164    2023-11-27 [1] CRAN (R 4.3.1)
```

```
##  nloptr            2.0.3      2022-05-26 [1] CRAN (R 4.3.0)
##  nnet              7.3-19     2023-05-03 [1] CRAN (R 4.3.1)
##  papaja          * 0.1.1.9001 2024-05-07 [1] Github (crsh/papaja@bd1aa4a)
##  parallelly        1.37.1     2024-02-29 [1] CRAN (R 4.3.1)
##  parameters        0.21.5     2024-02-07 [1] CRAN (R 4.3.1)
##  patchwork       * 1.2.0      2024-01-08 [1] CRAN (R 4.3.1)
##  phonR           * 1.0-7      2016-08-25 [1] CRAN (R 4.3.0)
##  pillar            1.9.0      2023-03-22 [1] CRAN (R 4.3.0)
##  pkgbuild          1.4.3      2023-12-10 [1] CRAN (R 4.3.1)
##  pkgconfig         2.0.3      2019-09-22 [1] CRAN (R 4.3.0)
##  pkgload           1.3.4      2024-01-16 [1] CRAN (R 4.3.1)
##  plyr              1.8.9      2023-10-02 [1] CRAN (R 4.3.1)
##  png               0.1-8      2022-11-29 [1] CRAN (R 4.3.0)
##  polyclip          1.10-6     2023-09-27 [1] CRAN (R 4.3.1)
##  posterior       * 1.5.0      2023-10-31 [1] CRAN (R 4.3.1)
##  prettyunits       1.2.0      2023-09-24 [1] CRAN (R 4.3.1)
##  profvis           0.3.8      2023-05-02 [1] CRAN (R 4.3.0)
##  progress          1.2.3      2023-12-06 [1] CRAN (R 4.3.1)
##  promises          1.2.1      2023-08-10 [1] CRAN (R 4.3.0)
##  proxy             0.4-27     2022-06-09 [1] CRAN (R 4.3.0)
##  purrr           * 1.0.2      2023-08-10 [1] CRAN (R 4.3.0)
##  QuickJSR          1.1.3      2024-01-31 [1] CRAN (R 4.3.1)
##  R6                2.5.1      2021-08-19 [1] CRAN (R 4.3.0)
##  rbibutils         2.2.16     2023-10-25 [1] CRAN (R 4.3.1)
##  RColorBrewer      1.1-3      2022-04-03 [1] CRAN (R 4.3.0)
##  Rcpp            * 1.0.12     2024-01-09 [1] CRAN (R 4.3.1)
##  RcppParallel      5.1.7      2023-02-27 [1] CRAN (R 4.3.0)
##  Rdpack            2.6        2023-11-08 [1] CRAN (R 4.3.1)
##  readr           * 2.1.5      2024-01-10 [1] CRAN (R 4.3.1)
##  remotes           2.4.2.1    2023-07-18 [1] CRAN (R 4.3.0)
```

```
##  reshape2          1.4.4     2020-04-09 [1] CRAN (R 4.3.0)
##  rlang           * 1.1.3     2024-01-10 [1] CRAN (R 4.3.1)
##  rmarkdown         2.26      2024-03-05 [1] CRAN (R 4.3.1)
##  rmutil            1.1.10    2022-10-27 [1] CRAN (R 4.3.0)
##  rpart             4.1.23    2023-12-05 [1] CRAN (R 4.3.1)
##  rprojroot         2.0.4     2023-11-05 [1] CRAN (R 4.3.1)
##  rsample         * 1.2.0     2023-08-23 [1] CRAN (R 4.3.0)
##  rstan             2.32.6    2024-03-05 [1] CRAN (R 4.3.1)
##  rstantools        2.4.0     2024-01-31 [1] CRAN (R 4.3.1)
##  rstudioapi        0.15.0    2023-07-07 [1] CRAN (R 4.3.0)
##  sandwich          3.1-0     2023-12-11 [1] CRAN (R 4.3.1)
##  scales            1.3.0     2023-11-28 [1] CRAN (R 4.3.1)
##  sessioninfo       1.2.2     2021-12-06 [1] CRAN (R 4.3.0)
##  sf                1.0-15    2023-12-18 [1] CRAN (R 4.3.1)
##  shiny             1.8.0     2023-11-17 [1] CRAN (R 4.3.1)
##  shinyjs           2.1.0     2021-12-23 [1] CRAN (R 4.3.0)
##  shinystan         2.6.0     2022-03-03 [1] CRAN (R 4.3.0)
##  shinythemes       1.2.0     2021-01-25 [1] CRAN (R 4.3.0)
##  spatial           7.3-17    2023-07-20 [1] CRAN (R 4.3.0)
##  stable            1.1.6     2022-03-02 [1] CRAN (R 4.3.0)
##  stabledist        0.7-1     2016-09-12 [1] CRAN (R 4.3.0)
##  StanHeaders       2.32.6    2024-03-01 [1] CRAN (R 4.3.1)
##  statip            0.2.3     2019-11-17 [1] CRAN (R 4.3.0)
##  stringi           1.8.4     2024-05-06 [1] CRAN (R 4.3.1)
##  stringr         * 1.5.1     2023-11-14 [1] CRAN (R 4.3.1)
##  supunsup        * 0.2.0     2023-06-26 [1] Github (kleinschmidt/phonetic-sup-unsup@5c5117
##  survival          3.5-8     2024-02-14 [1] CRAN (R 4.3.1)
##  svglite           2.1.3     2023-12-08 [1] CRAN (R 4.3.1)
##  svUnit            1.0.6     2021-04-19 [1] CRAN (R 4.3.0)
##  systemfonts       1.0.6     2024-03-07 [1] CRAN (R 4.3.1)
```

```
##  tensorA           0.36.2.1   2023-12-13 [1] CRAN (R 4.3.1)
##  terra             1.7-71     2024-01-31 [1] CRAN (R 4.3.1)
##  TH.data           1.1-2      2023-04-17 [1] CRAN (R 4.3.0)
##  threejs           0.3.3      2020-01-21 [1] CRAN (R 4.3.0)
##  tibble          * 3.2.1      2023-03-20 [1] CRAN (R 4.3.0)
##  tidybayes       * 3.0.6      2023-08-12 [1] CRAN (R 4.3.0)
##  tidyr           * 1.3.1      2024-01-24 [1] CRAN (R 4.3.1)
##  tidyselect        1.2.1      2024-03-11 [1] CRAN (R 4.3.1)
##  tidyverse       * 2.0.0      2023-02-22 [1] CRAN (R 4.3.0)
##  timechange        0.3.0      2024-01-18 [1] CRAN (R 4.3.1)
##  timeDate          4032.109   2023-12-14 [1] CRAN (R 4.3.1)
##  timeSeries        4032.109   2024-01-14 [1] CRAN (R 4.3.1)
##  tinylabels      * 0.2.4      2023-09-02 [1] CRAN (R 4.3.0)
##  transformr        0.1.5      2024-02-26 [1] CRAN (R 4.3.1)
##  tweenr            2.0.3      2024-02-26 [1] CRAN (R 4.3.1)
##  tzdb              0.4.0      2023-05-12 [1] CRAN (R 4.3.0)
##  units             0.8-5      2023-11-28 [1] CRAN (R 4.3.1)
##  urlchecker        1.0.1      2021-11-30 [1] CRAN (R 4.3.0)
##  usethis           2.2.3      2024-02-19 [1] CRAN (R 4.3.1)
##  utf8              1.2.4      2023-10-22 [1] CRAN (R 4.3.1)
##  V8                4.4.2      2024-02-15 [1] CRAN (R 4.3.1)
##  vctrs             0.6.5      2023-12-01 [1] CRAN (R 4.3.1)
##  viridis           0.6.5      2024-01-29 [1] CRAN (R 4.3.1)
##  viridisLite       0.4.2      2023-05-02 [1] CRAN (R 4.3.0)
##  vroom             1.6.5      2023-12-05 [1] CRAN (R 4.3.1)
##  webshot         * 0.5.5      2023-06-26 [1] CRAN (R 4.3.0)
##  withr             3.0.0      2024-01-16 [1] CRAN (R 4.3.1)
##  xfun              0.43       2024-03-25 [1] CRAN (R 4.3.1)
##  xml2              1.3.6      2023-12-04 [1] CRAN (R 4.3.1)
##  xtable            1.8-4      2019-04-21 [1] CRAN (R 4.3.0)
```

```
##  xts               0.13.2     2024-01-21 [1] CRAN (R 4.3.1)
##  yaml              2.3.8      2023-12-11 [1] CRAN (R 4.3.1)
##  zoo               1.8-12     2023-04-13 [1] CRAN (R 4.3.0)
##
##  [1] /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/library
##
## ------------------------------------------------------------------------------------------
```

Table S13

*Have participants converged against the PSE expected from idealized learner? This table compares changes in participants' categorization function against those expected from idealized learners, complementing Table* **??**.

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Test block 1** | | | | | |
| $\lvert\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{baseline}})\rvert > 0$ | 0.41 | 0.08 | [0.23, 0.58] | $\geq 8000$ | 1.00 |
| $\lvert\Delta(PSE_{ideal_{+10}}, PSE_{actual_{+10}})\rvert > 0$ | 0.22 | 0.08 | [0.05, 0.37] | $\geq 8000$ | 1.00 |
| $\lvert\Delta(PSE_{ideal_{+40}}, PSE_{actual_{+40}})\rvert > 0$ | 0.47 | 0.07 | [0.35, 0.62] | $\geq 8000$ | 1.00 |
| **Test block 2** | | | | | |
| $\lvert\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{baseline}})\rvert > 0$ | 0.33 | 0.06 | [0.21, 0.44] | $\geq 8000$ | 1.00 |
| $\lvert\Delta(PSE_{ideal_{+10}}, PSE_{actual_{+10}})\rvert > 0$ | 0.17 | 0.05 | [0.05, 0.27] | $\geq 8000$ | 1.00 |
| $\lvert\Delta(PSE_{ideal_{+40}}, PSE_{actual_{+40}})\rvert > 0$ | 0.32 | 0.07 | [0.19, 0.49] | $\geq 8000$ | 1.00 |
| **Test block 3** | | | | | |
| $\lvert\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{baseline}})\rvert > 0$ | 0.35 | 0.04 | [0.25, 0.43] | $\geq 8000$ | 1.00 |
| $\lvert\Delta(PSE_{ideal_{+10}}, PSE_{actual_{+10}})\rvert > 0$ | 0.16 | 0.06 | [0.05, 0.27] | $\geq 8000$ | 1.00 |
| $\lvert\Delta(PSE_{ideal_{+40}}, PSE_{actual_{+40}})\rvert > 0$ | 0.27 | 0.06 | [0.15, 0.41] | $\geq 8000$ | 1.00 |
| **Test block 4** | | | | | |
| $\lvert\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{baseline}})\rvert > 0$ | 0.32 | 0.06 | [0.2, 0.44] | $\geq 8000$ | 1.00 |
| $\lvert\Delta(PSE_{ideal_{+10}}, PSE_{actual_{+10}})\rvert > 0$ | 0.15 | 0.04 | [0.05, 0.23] | $\geq 8000$ | 1.00 |
| $\lvert\Delta(PSE_{ideal_{+40}}, PSE_{actual_{+40}})\rvert > 0$ | 0.29 | 0.05 | [0.2, 0.46] | $\geq 8000$ | 1.00 |
| **Does convergence against ideal PSE differ across conditions (asymmetric shrinkage)?** | | | | | |
| **Test block 1** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+10}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | 0.00 | 0.00 | [0, 0] | 0 | 0.00 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+40}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | 0.00 | 0.00 | [0, 0] | 0 | 0.00 |
| **Test block 2** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+10}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | 0.06 | 0.54 | [-1.87, 1.61] | 1.2 | 0.54 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+40}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | 0.14 | 0.33 | [-0.47, 0.96] | 2 | 0.67 |
| **Test block 3** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+10}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | 0.09 | 0.53 | [-1.87, 1.65] | 1.3 | 0.57 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+40}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | 0.28 | 0.30 | [-0.28, 1.07] | 5 | 0.83 |
| **Test block 4** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+10}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | 0.12 | 0.48 | [-1.74, 1.6] | 1.5 | 0.60 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+40}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | 0.16 | 0.32 | [-0.46, 0.92] | 2.3 | 0.70 |

Table S14
*Comparison of actual changes in participants' categorization function against those expected from idealized learners. Both participants' and the ideal learner's PSEs are compared against the median PSE of the five cross-validated idealized pre-exposure listeners.*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Test block 1** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{+10}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | -0.08 | 0.48 | [-0.91, 0.95] | 0.7 | 0.43 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{predicted_{prior}})}{\Delta(PSE_{ideal_{+40}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | -0.07 | 0.27 | [-0.61, 0.43] | 0.6 | 0.39 |
| **Test block 2** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{+10}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | -0.05 | 0.31 | [-0.59, 0.6] | 0.8 | 0.44 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{+40}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | 0.08 | 0.22 | [-0.39, 0.48] | 1.7 | 0.63 |
| **Test block 3** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{+10}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | 0.01 | 0.30 | [-0.54, 0.63] | 1.0 | 0.51 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{+40}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | 0.22 | 0.18 | [-0.16, 0.57] | 6.1 | 0.86 |
| **Test block 4** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{+10}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | 0.02 | 0.26 | [-0.46, 0.57] | 1.2 | 0.54 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{+40}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | 0.10 | 0.20 | [-0.34, 0.46] | 2.3 | 0.69 |

Table S15
*Comparison of actual changes in participants' categorization function against those expected from idealized learners. Both participants' and the ideal learner's PSEs are compared against the median PSE of the five cross-validated idealized pre-exposure listeners.*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Exposure block 1** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{+10}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | -0.03 | 0.21 | [-0.42, 0.47] | 0.8 | 0.45 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{+40}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | -0.14 | 0.18 | [-0.56, 0.16] | 0.3 | 0.21 |
| **Exposure block 2** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{+10}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | 0.02 | 0.20 | [-0.32, 0.47] | 1.2 | 0.54 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{+40}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | -0.06 | 0.15 | [-0.48, 0.18] | 0.5 | 0.34 |
| **Exposure block 3** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{+10}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | 0.08 | 0.17 | [-0.21, 0.46] | 2.3 | 0.70 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{+40}}, PSE_{CV_{prior}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{CV_{prior}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{CV_{prior}})}$ | -0.03 | 0.13 | [-0.32, 0.27] | 0.7 | 0.41 |

Table S16

*Comparison of actual changes in participants' categorization function against those expected from idealized learners. Both participants' and the ideal learner's PSEs are compared against the PSE of participants during Test 1 (averaging across all three conditions).*

| Hypothesis | Est. | SE | 90%-CI | BF | $p_{post}$ |
|---|---|---|---|---|---|
| **Test block 1** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+10}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | -0.11 | 0.50 | [-1.13, 1.05] | 0.7 | 0.41 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+40}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | -0.04 | 0.17 | [-0.37, 0.31] | 0.7 | 0.41 |
| **Test block 2** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+10}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | -0.09 | 0.38 | [-1.06, 0.62] | 0.7 | 0.41 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+40}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | 0.11 | 0.30 | [-0.45, 0.67] | 1.9 | 0.66 |
| **Test block 3** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+10}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | -0.02 | 0.36 | [-0.99, 0.65] | 0.9 | 0.47 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+40}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | 0.26 | 0.26 | [-0.28, 0.79] | 4.9 | 0.83 |
| **Test block 4** | | | | | |
| $\frac{\Delta(PSE_{actual_{+10}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+10}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | -0.01 | 0.31 | [-0.87, 0.58] | 0.9 | 0.49 |
| $\frac{\Delta(PSE_{actual_{+40}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{+40}}, PSE_{actual_{pre}})} > \frac{\Delta(PSE_{actual_{baseline}}, PSE_{actual_{pre}})}{\Delta(PSE_{ideal_{baseline}}, PSE_{actual_{pre}})}$ | 0.13 | 0.28 | [-0.44, 0.67] | 2.2 | 0.69 |