Listeners adjust their prior expectations as they adapt to speech of an unfamiliar talker

Maryann Tan[1,2], Maryann Tan[2,3], & T F Jaeger[2]

[1] Centre for Research on Bilingualism, University of Stockholm

[2] Brain and Cognitive Sciences, University of Rochester

[3] Computer Science, University of Rochester

Author Note

We are grateful to ### ommitted for review ###

Correspondence concerning this article should be addressed to Maryann Tan, YOUR ADDRESS. E-mail: maryann.tan@biling.su.se

Abstract

YOUR ABSTRACT GOES HERE. All data and code for this study are shared via OSF, including the R markdown document that this article is generated from, and an R library that implements the models we present.

Keywords: speech perception; perceptual adaptation; distributional learning; ...

Word count: X

Listeners adjust their prior expectations as they adapt to speech of an unfamiliar talker

TO-DO

0.1 Highest priority

- MARYANN

0.1.1 Priority

- FLORIAN

0.2 To do later

- Everyone: Eat ice-cream and perhaps have a beer.

1 Introduction

Talkers who share a common language vary in the way they pronounce its linguistic categories. Yet, listeners of the same language background typically cope with such variation without much effort. In scenarios where a talker produces those categories in an unexpected and unfamiliar way, comprehending their speech may pose a real challenge. However, brief exposure to the talker's accent (sometimes just minutes) can be sufficient for the listener to overcome any initial comprehension difficulty (e.g. Bradlow & Bent, 2008; Clarke & Garrett, 2004; X. Xie, Liu, & Jaeger, 2021; X. Xie et al., 2018). This adaptive skill is in a sense, trivial for any expert language user but becomes complex when considered from the angle of acoustic-cue-to-linguistic-category mappings. Since talkers differ in countless ways and each listening occasion is different in circumstance, there is not a single set of cues that can be definitively mapped to each linguistic category. Listeners instead have to contend with many possible cue-to-category mappings and infer the intended category of the talker. How listeners achieve prompt and accurate comprehension of speech in spite of this variability (the classic "lack of invariance" problem) remains the overarching aim of speech perception research.

Researchers have been exploring the hypothesis that listeners solve this perceptual problem by exploiting their knowledge gained from experience with different talkers. This knowledge is often implicit and context contingent since listeners are sensitive to both social and environmental cues (e.g. age, sex, group identity, native language etc.) that are relevant for optimal speech perception. Impressively, shifts in perception can be induced implicitly through subtle cues such as the presence of cultural artefacts that hint at talker provenance, (Hay & Drager, 2010) and explicitly such as when the listener is instructed to imagine a talker as a man or a woman (Johnson, Strand, & D'Imperio, 1999). While these and other related effects of exposure-induced changes speak to the malleability of human perception, it remains unclear how human perceptual systems strike the balance between stability and flexibility.

One possibility is that listeners continuously update their implicit knowledge with each talker encounter by integrating prior knowledge of cue-to-category distributions with the statistics of the current talker's productions, leading to changes in representations which affect listener categorisation behaviour. Broadly speaking, many theoretical accounts would agree with this assertion. Connectionist (McClelland & Elman 1986; Luce & Pisoni, 1998), and Bayesian models of spoken word recognition (Norris & McQueen, 2008) and adaptation (Kleinschmidt & Jaeger, 2015) are generative systems that abstract the frequency of input. Even exemplar models of speech perception (Goldinger 1996, 1998; Johnson, 1997; Pierrehumbert 2001) which encode high fidelity memories of speaker-specific phonetic detail converge to a level of generalisation due to effects of token frequency (**Pierrehumbert2003?**; **DragerKirtley2016?**).

At the level of acoustic-phonetic input, listeners' implicit knowledge refer to the way relevant acoustic cues that distinguish phonological categories are distributed across talkers within a linguistic system. Talkers of US-English, for instance, distinguish the /d/-/t/ contrasts primarily through the voice-onset-time (VOT) acoustic cue. Given its relevance for telling word pairs such as "din" and "tin" apart, a distributional learning hypothesis would posit that listeners learn the distribution of VOT cues when talkers produce those stop consonant contrasts in word contexts. Earliest evidence for listener sensitivity to individual talker statistics in the domain of stop consonants come from studies such as Allen & Miller (2004, also Theodore & Miller, 2010) but more recent studies that formalise the problem of speech perception as rational inference have shown that listeners' behavioural responses are probabilistic function of the exposure talker's statistics (Clayards, Tanenhaus, Aslin, & Jacobs, 2008a; Kleinschmidt & Jaeger, 2016; and Theodore & Monto, 2019).

Clayards et al. (2008a) for instance found that listeners responded with greater uncertainty after they were exposed to VOT distributions for a "beach-peach" contrast that had wider variances as compared to another group who had heard the same contrasts with narrower variances. Across both wide and narrow conditions, the mean values of the voiced and voiceless categories were kept constant and set at values that were close to the expected means for /b/ and /p/ in US English. The study was one of the first to demonstrate that at least in the context of an experiment, listeners categorisation behaviour was a function of the variance of the exposure talker's cue distributions – listeners who were exposed to a wide distribution of VOTs showed greater uncertainty in their perception of the stimuli, exhibiting a flatter categorisation function on average, compared to listeners who were exposed to a narrow distribution.

In a later study Kleinschmidt and Jaeger (2016) tested listener response to talker statistics by shifting the means of the voiced and voiceless categories between conditions. Specifically, the mean values for /b/ and /p/ were shifted rightwards in varying durations, as well as leftwards, from the expected mean values of a typical American English talker while the category variances remained identical and the distance between the category means were kept constant. With this manipulation of means they were able to investigate how inclined listeners are to adapt their categorisation behaviors when the statistics of the exposure talker were shifted beyond the bounds of a typical talker.

In all exposure conditions, listeners on average adapted to the exposure talker by shifting their categorization towards the boundary implied by the exposure distribution. However, in all conditions, listener categorization fell short of the predicted ideal categorization boundary. This difference between the observed and predicted categorization functions was larger, the greater the magnitude of the shift from the typical talker's distribution, suggesting adaptation was constrained by listeners' prior experience.

The study we report here builds on the pioneering work of Clayards et al. (2008a) and Kleinschmidt and Jaeger (2016) with the aim to shed more light on the role of prior implicit knowledge on adaptation to an unfamiliar talker.

Specifically, while K&J16 demonstrated how prior beliefs of listeners can be inferred computationally from post-exposure categorisation, their experiment was not designed to capture listener categorisation data before exposure to a novel talker. Nor did they run intermittent tests to scrutinise the progress of adaptation. In the ideal adapter framework, listener expectations are predicted to be rationally updated through integration with the incoming speech input and thus can theoretically be analysed on a trial-by-trial basis. The overall design of the studies reported here were motivated by our aim to understand this incremental belief-updating process which has not been closely studied in previous work. We thus address the limitations of previous work and in conjunction, make use of ideal observer models to validate baseline assumptions that accompany this kind of speech perception study – that listeners hold prior expectations or beliefs about cue distributions based on previously experienced speech input (here taken to mean native AE listeners' lifetime of experience with AE). Arriving at a definitive conclusion of what shape and form those beliefs take is beyond the scope of this study however we attempt to explore the various proposals that have emerged from more than half a century of speech perception research.

A secondary aim was to begin to address possible concerns of ecological validity of prior work. While no speech stimuli is ever ideal, previous work on which the current study is based did have limitations in one or two aspects:the artificiality of the stimuli or the artificiality of the distributions. For e.g. (Clayards et al., 2008a) and (Kleinschmidt & Jaeger, 2016) made use of synthesised stimuli that were robotic or did not sound human-like. The second way that those studies were limited was that the exposure distributions of the linguistic categories had identical variances (see also Theodore & Monto, 2019) unlike what is found in production data where the variance of the voiceless categories are typically wider than that of the voiced category (Chodroff & Wilson, 2017). We take modest steps to begin to improve the ecological validity of this study while balancing the need for control through lab experiments by employing more natural sounding stimuli as well as by setting the variances of our exposure distributions to better reflect empirical data on production (see section x.xx. of SI).

2 Experiment 1: Listener's expectations prior to informative exposure

Experiment 1 investigates native (L1) US English listeners' categorization of word-initial stop voicing by an unfamiliar female L1 US English talker, prior to more informative exposure. Specifically, listeners heard isolated recordings from a /d/-/t/ continuum, and had to respond which word they heard (e.g., "din" or "tin"). The recordings varied in voice onset time (VOT), the primary phonetic cue to word-initial stop voicing in L1 US English, as well as correlated secondary cues (f0 and rhyme duration). Critically, exposure was relatively uninformative about the talker's use of the phonetic cues in that all phonetic realizations occurred equally often. The design of Experiment 1 serves two goals.

The first goal is methodological. We use Experiment 1 to test basic assumptions about the paradigm and stimuli we employ in the remainder of this study. We obtain estimates of the category boundary between /d/ and /t/ *for the specific stimuli used in Experiment 2*, as perceived by *the type of listeners we seek to recruit for Experiment 2*. We also test whether prolonged testing across the phonetic continuum changes listeners' categorization behavior. Previous work has found that prolonged testing on uniform distributions can reduce the effects of previous exposure (Liu & Jaeger, 2018a; e.g., **mitterer2011?**), at least in listeners of the age group we recruit from (Scharenborg & Janse, 2013). However, these studies employed only a small number of 5-7 perceptually highly ambiguous stimuli, each repeated many times. In Experiment 1, we employ a much larger set of stimuli that span the entire continuum from very clear /d/s to very clear /t/s, each presented only twice. If prolonged testing changes listeners' responses, this has to be taken into account in the design of Experiment 2.

The second purpose of Experiment 1 is to introduce and illustrate relevant theory. We compare different models of listeners' prior expectations against listeners' categorization responses in Experiment 1. The different models all aim to capture the implicit expectations of an L1 adult listener of US English might have about the mapping from acoustic cues to /d/ and /t/ based on previously experienced speech input. As we describe in more detail after the presentation of the experiment, the models differ, however, in whether these prior expectations take into account that talkers can differ in the way they realize /d/ and /t/. This ability to take into account talker differences even prior to more informative exposure is predicted—though through qualitatively different mechanisms, as we discuss below—both by normalization accounts (Cole, Linebaugh, Munson, & McMurray, 2010; McMurray & Jongman, 2011) and by accounts that attribute adaptive speech perception to changes in category representations (Bayesian ideal adaptor theory, Kleinschmidt & Jaeger, 2015; EARSHOT, Magnuson et al., 2020; episodic theory, Goldinger, 1998; exemplar theory, Johnson, 1997; Pierrehumbert, 2001). It is, however, unexpected under accounts that attribute adaptive speech perception solely to ad-hoc changes in decision-making. We did not expect that Experiment 1 yields a decisive conclusion with regard to this second goal, which is also addressed in Experiment 2. Rather, we use Experiment 1 as a presentationally convenient way to introduce some of the different models and provide readers with initial intuitions about what experiments of this type can and cannot achieve.

2.1 Methods

2.1.1 Participants

Participants were recruited over Amazon's Mechanical Turk platform, and paid $2.50 each (for a targeted remuneration of $6/hour). The experiment was only visible to Mechanical Turk participants who (1) had an IP address in the United States, (2) had an approval rating of 95% based on at least 50 previous assignments, and (3) had not previously participated in any experiment on stop voicing from our lab.

24 L1 US English listeners (female = 9; mean age = 36.2 years; SD age = 9.2 years) completed the experiment. To be eligible, participants had to confirm that they (1) spent at least the first 10 years of their life in the US speaking only English, (2) were in a quiet place, and (3) wore in-ear or over-the-ears headphones that cost at least $15.

2.1.2 Materials

We recorded multiple tokens of four minimal word pairs ("dill"/"till", "dim"/"tim", "din"/"tin", and "dip"/"tip") from a 23-year-old, female L1 US English talker with a mid-Western accent. These recordings were used to create four natural-sounding minimal pair VOT continua (dill-till, dip-tip, din-tin, and dip-tip) using a Praat script (Winn, 2020). The full procedure is described in the supplementary information (SI, **??**). The VOT continua ranged from -100ms VOT to +130ms VOT in 5ms steps. Experiment 1 employs 24 of these steps (-100, -50, -10, 5 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 100, 110, 120, 130). VOT tokens in the lower and upper ends were distributed over larger increments because stimuli in those ranges were expected to elicit floor and ceiling effects, respectively.

We further set the F0 at vowel onset to follow the speaker's natural correlation which was estimated through a linear regression analysis of all the recorded speech tokens. We did this so that we could determine the approximate corresponding f0 values at each VOT value along the continua as predicted by this talker's VOT. The duration of the vowel was set to follow the natural trade-off relation with VOT reported in Allen and Miller (1999). This approach closely resembles that taken in Theodore and Monto (2019), and resulted in continuum steps that sound highly natural (unlike the robotic-sounding stimuli employed in Clayards et al., 2008a; Kleinschmidt & Jaeger, 2016). All stimuli are available as part of the OSF repository for this article.

In addition to the critical minimal pair continua we also recorded three words that did not did not contain any stop consonant sounds ("flare", "share", and "rare"). These word recordings were used as catch trials. Stimulus intensity was set to 70 dB sound pressure level for all recordings.

2.1.3 Procedure

The code for the experiment is available as part of the OSF repository for this article. A live version is available at (https://www.hlp.rochester.edu//experiments/DLVOT/series-A/experiment-A.html?list_test=NORM-A-forward-test). The first page of the experiment informed participants of their rights and the requirements for the experiment: that they had to be native listeners of English, wear headphones for the entire duration of the experiment, and be in a quiet room without distractions. Participants had to pass a headphone test, and were asked to keep the volume unchanged throughout the experiment. Participants could only advance to the start of the experiment by acknowledging each requirement and consenting to the guidelines of the Research Subjects Review Board of the University of Rochester.

On the next page, participants were informed about the task for the remainder of the experiment. They were informed that they would heard a female talker speak a single word on each trial, and had to select which word they heard. Participants were instructed to listen carefully and answer as quickly and as accurately as possible. They were also alerted to the fact that the recordings were subtly different and therefore may sound repetitive. This was done to encourage their full attention.

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, as shown in Figure 1. At 1000ms from trial onset, the fixation dot would turn bright green and an audio recording from the matching minimal pair continuum started playing. Participants were required to click on the word they heard. For each participant, /d/-initial words were either always displayed on the left side or always displayed on the right side. Across participants, this ordering was counter-balanced. After participants clicked on the word, the next trial began.

Participants heard 192 target trials (four minimal pair continua, each with 24 VOT steps,

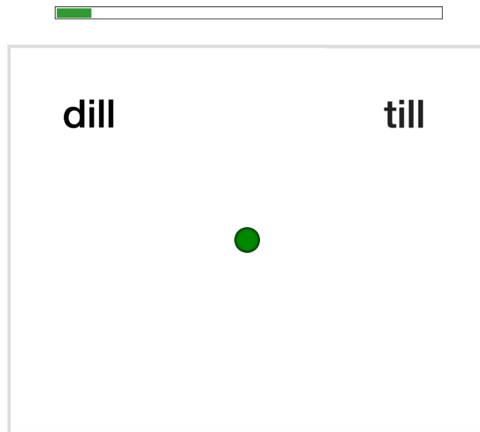

Figure 1. Example trial display. The words were displayed 500ms after trial onset and the audio recording of the word was played 1000ms after trial onset

each heard twice). In addition, participants heard 12 catch trials. On catch trials, participant saw two written catch stimuli on the screen (e.g., “flare” and “rare”), and heard one of them (e.g. “rare”). Since these recordings were easily distinguishable, they served as a check on participant attention throughout the experiment.

The order of trials was randomized for each participant with the only constraint that no stimulus was repeated before each stimulus had been heard at least once. Catch trials were distributed randomly throughout the experiment with the constraint that no more than two catch trials would occur in a row. Participants were given the opportunity to take breaks after every 60 trials. Participants took an average of 12 minutes (SD = 4.8) to complete the 204 trials, after which they answered a short survey about the experiment.

2.1.4 Exclusions

We excluded from analysis participants who committed more than 2 errors out of the 12 catch trials (<83% accuracy, N = 3), participants with an average reaction time (RT) more than three standard deviations from the mean of the by-participant means (N = 0), and participants who reported not to have used headphones (N = 0) or not to be native (L1) speakers of US English (N = 0). For the remaining participants, trials that were more than three SDs from the participant’s mean RT were excluded from analysis (1.6%). Finally, we excluded participants (N = 0) who had less than 50% data remaining after these exclusions.

2.2 Behavioral results

We first present the behavioral analyses of participants' categorisation responses. Then we compare participants' responses to the predictions of different models fit on the distribution of stop voicing cues in a large database of L1 US English productions of word-initial /d/s and /t/s (Chodroff & Wilson, 2018).

2.2.1 Analysis approach

The goal of our behavioral analyses was to address three methodological questions that are of relevance to Experiment 2: (1) whether our stimuli resulted in 'reasonable' categorisation functions, (2) whether these functions differed between the four minimal pair items, and (3) whether participants' categorisation functions changed throughout the 192 test trials.

To address these questions, we fit a single Bayesian mixed-effects psychometric model to participants' categorization responses on critical trials (e.g., **prins2011?**). This model is essentially an extension of mixed-effects logistic regression that also takes into account attentional lapses. A failure to do so—while commonplace in research on speech perception (incl. our own work, but see Clayards, Tanenhaus, Aslin, & Jacobs, 2008b; Kleinschmidt & Jaeger, 2016)—can lead to biased estimates of categorization boundaries (e.g., Wichmann & Hill, 2001). The mixed-effects psychometric model describes the probability of "t"-responses as a weighted mixture of a lapsing-model and a perceptual model. The lapsing model is a mixed-effects logistic regression (Jaeger, 2008) that predicts participant responses that are made independent of the stimulus—for example, responses that result from attentional lapses. These responses are independent of the stimulus, and depend only on participants' response bias. The perceptual model is a mixed-effects logistic regression that predicts all other responses, and captures stimulus-dependent aspects of participants' responses. The relative weight of the two models is determined by the lapse rate, which is described by a third mixed-effects logistic regression.

The *lapsing model* only contained an intercept (the response bias in log-odds) and by-participant random intercepts. Similarly, the *model for the lapse rate* only had an intercept (the lapse rate) and by-participants random intercepts. No by-item random effects were included for the lapse rate nor lapsing model since these parts of the analysis—by definition—describe stimulus-*in*dependent behavior. The *perceptual model* included an intercept and VOT, as well as the full random effect structure by participants and items (the four minimal pair continua), including random intercepts and random slopes by participant and minimal pair. We did not model the random effects of trial to reduce model complexity. This potentially makes our analysis of trials in the model anti-conservative. Finally, the models included the covariance between by-participant random effects across the three linear predictors for the lapsing model, lapse rate model, and perceptual model. This allows us to capture whether participants who lapse more often have, for example, different response biases or different sensitivity to VOT (after accounting for lapsing).

We fit the model using the package `brms` (Bürkner, 2017) in R (R Core Team, 2021a; RStudio Team, 2020). Following previous work from our lab (Hörberg & Jaeger, 2021; X. Xie et al., 2021), we used weakly regularizing priors to facilitate model convergence. For fixed effect parameters, we standardized continuous predictors (VOT) by dividing through twice their standard deviation (Gelman, 2008), and used Student priors centered around zero with a scale of 2.5 units (following Gelman, Jakulin, Pittau, & Su, 2008) and 3 degrees of freedom. 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 (**Lewandowski2009?**). Four chains with 2000 warm-up samples and 2000 posterior samples each were fit. No divergent transitions after warm-up were observed, and all $\hat{R}$ were close to 1.

2.2.2 Expectations

Based on previous experiments, we expected a strong positive effect of VOT, with increasing proportions of "t"-responses for increasing VOTs. We did not have clear expectations for the effect of trial other than that responses should become more uniformed (i.e move towards 50-50 "d"/"t"-bias or 0-log-odds) as the experiment progressed (Liu & Jaeger, 2018b) due to the un-informativeness of the stimuli. Previous studies with similar paradigms have typically found lapse rates of 0-10% (< -2.2 log-odds, e.g., Clayards et al., 2008a; Kleinschmidt & Jaeger, 2016).

The lapse rate was estimated to be on the slightly larger side, but within the expected

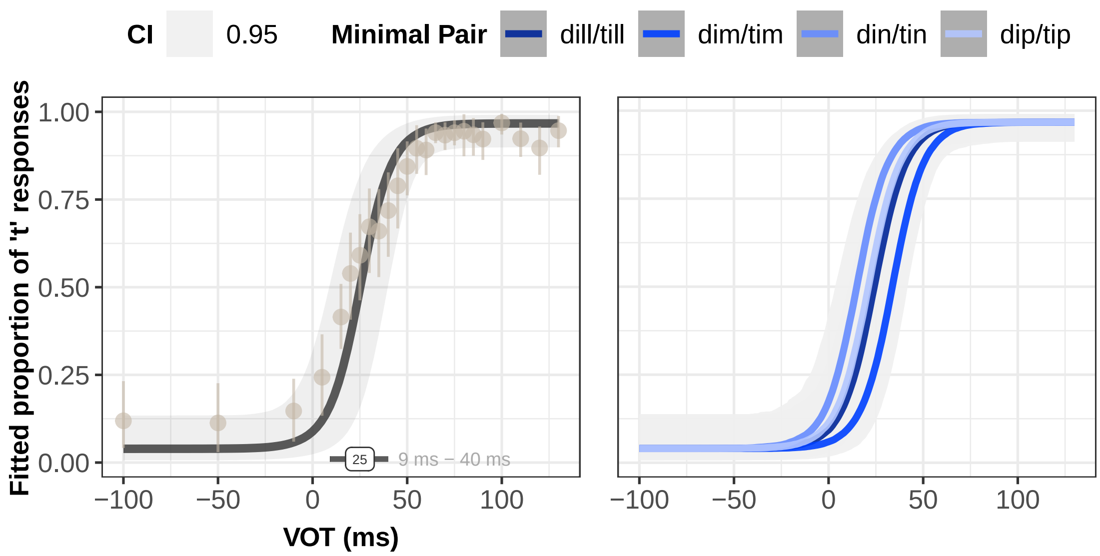

Figure 2. Categorisation functions and points of subjective equality (PSE) derived from the Bayesian mixed-effects psychometric model fit to listeners' responses in Experiment 1. The categorization functions include lapse rates and biases. The PSEs correct for lapse rates and lapse biases (i.e., they are the PSEs of the perceptual component of the psychometric model).[1] **Left:** Effects of VOT, lapse rate, and lapse bias, while marginalizing over trial effects as well as all random effects. Vertical point ranges represent the mean proportion and 95% bootstrapped CIs of participants' "t"-responses at each VOT step. Horizontal point ranges denote the mean and 95% quantile interval of the points of subjective equality (PSE), derived from the 8000 posterior samples of the population parameters. **Right:** The same but showing the fitted categorization functions for each of the four minimal pair continua. Participants' responses are omitted to avoid clutter.

range (7.5 %, 95%-CI: 2.2 to 21.2%; Bayes factor: 1,599 90%-CI : -3.54 to -1.53). Maximum a posteriori (MAP) estimates of by-participant lapse rates ranged from XX . Very high lapse rates were estimated for four of the participants with one in particular whose CI indicated exceptionally high uncertainty. These lapse rates might reflect data quality issues with Mechanical Turk that started to emerge over recent years (see **REFS?**; and, specifically for experiments on speech perception, **cummings2023?**), and we return to this issue in Experiment 2.

The response bias were estimated to slightly favor "t"-responses (53.4 %, 95%-CI: 17.1 to 82.1%; Bayes factor: 1.52 90%-CI : -1.21 to 1.31), as also visible in Figure 2 (left). Unsurprisingly, the psychometric model suggests high uncertainty about the participant-specific response biases, as it is difficult to reliably estimate participant-specific biases while also accounting for trial and VOT effects (range of by-participant MAP estimates: XX). For all but four participants, the 95% CI includes the hypothesis that responses were unbiased. Of the remaining four participants, three were biased towards "t"-responses and one was biased toward "d"-responses.

There was no convincing evidence of a main effect of trial ($\hat{\beta}$ = -0.2 95%-CI: -0.6 to 0.4; Bayes factor: 2.71 90%-CI : -0.57 to 0.26). Given the slight overall bias towards "t"-responses, the direction of this effect indicates that participants converged towards a 50/50 bias as the test phase proceeded. This is also evident in Figure 2 (right). In contrast, there was clear evidence for a positive main effect of VOT on the proportion of "t"-responses ($\hat{\beta}$ = 12.6 95%-CI: 9.8 to 15.5; Bayes factor: Inf 90%-CI : 10.27 to 15.04). The effect of VOT was consistent across all minimal pair words as evident from the slopes of the fitted lines by minimal pair 2 (left). MAP estimates of by minimal pair slopes ranged from . The by minimal-pair intercepts were more varied (MAP estimates:) with one of the pairs, dim/tim having a slightly lower intercept resulting in fewer 't'-responses on average. In all, this justifies our assumptions that word pair would not have a substantial effect on categorisation behaviour. From the parameter estimates of the overall fit we obtained the category boundary from the point of subjective equality (PSE) `r( descale(-(summary(fit_mix)$fixed["mu2_Intercept", 1] / summary(fit_mix)$fixed["mu2_sVOT", 1]), VOT.mean_exp1, VOT.sd_exp1) ms)` which we use for the design of Experiment 2.

Finally to accomplish the first goal of experiment 1, we look at the interaction between VOT and trial. There was weak evidence that the effect of VOT decreased across trials ($\hat{\beta}$ = -0.6 95%-CI: -2.6 to 1.4; Bayes factor: 2.76 90%-CI : -2.27 to 1.05). The direction of this change—towards more shallow VOT slopes as the experiment progressed—makes sense since the test stimuli were not informative about the talker's pronunciation. Similar changes throughout prolonged testing have been reported in previous work. (Liu & Jaeger, 2018a, 2019; **REFS?**).

Overall, there was little evidence that participants substantially changed their categorisation behaviour as the experiment progressed. Still, to err on the cautious side, Experiment 2 employs shorter test phases.

2.3 Comparisons to model of adaptive speech perception

We now turn to final aim of experiment 1 which is to make use of computational models to begin to understand the implicit expectations that listeners hold when perceiving input that is uninformative of a talker's cue-to-category-mappings.

Speakers' productions can act as a proxy for listeners' implicit knowledge of the distributional patterns of cues. This production-perception relationship within a phonological system was observed in early work by (Abramson & Lisker, 1973) who found that production statistics of talkers along VOT aligned well with data from listeners who had categorised a separate set of synthesised VOT stimuli. This allows for the use of analytic models as tools for predicting categorisation behaviour from speech production data (Nearey & Hogan, 1986).

We apply this principle when fitting ideal observer (IO) models by linking the distributional patterns of talker productions to the categorisation behaviour of listeners. All models were trained on cue measurements extracted from an annotated database of 92 L1 US-English talkers' productions (Chodroff & Wilson, 2017) of word initial /d/ and /t/. By using IOs trained solely on production data to predict behaviour we avoid additional computational degrees of freedom and limit the risk of overfitting the model to the data.

The IOs' predictions apply Bayes' theorem to achieve optimal categorization; the posterior probability of recognising a token as the "t" category is function of its prior prior probability $p(c = t)$ and the probability of observing the token under the hypothesis that the talker intended the voiceless category $p(cue|c = t)$ taken as a proportion of the sum of probabilities of observing the token under all possible hypotheses.

We compare listener categorisation behaviour against the predictions of five IO models which reflect different assumptions about perceptual processes and the normalization (or lack thereof) of input. Beginning with a minimal model (raw VOT cues with no added perceptual noise), each successive model increased in complexity either with the addition the F0 cue or an assumption about speech encoding (Figure 4). All IO models were adjusted by the estimated lapse-rate from the psychometric fit to the perceptual data while bias was held at .5. In models that included perceptual noise we added a noise variance of 80ms (cf. Kronrod, Coppess, & Feldman, 2016) to the likelihoods. In addition to transforming the F0 cue measurements from raw Hz into Mel (Stevens & Volkmann, 1940) to reflect the tonotopy of the auditory system, normalization was applied to cues to compare effects of hypothesised pre-linguistic processes. We applied C-CuRE (McMurray and Jongman (2011); Toscano and McMurray (2015)), a general purpose normalization procedure which captures the hypothesis that listeners overcome multiple sources of variability by interpreting cues relative to the expected distribution of cues given the present context. While C-CuRE has the potential to be applied in various ways depending on the context to be evaluated, we implemented it in its most basic form, which is to center the cues– here VOT and F0 – relative to the talker population means across categories. In the final model we extended this centering process to the cues in the exposure stimuli. This additional step fully implements the assumption of pre-linguistic normalization being an automatic process.

Each of these models are then assessed for their goodness-of-fit to the categorisation data by comparing the likelihood of human responses under the assumptions represented by the respective IO models (Figure 4). For this we applied Luce's choice axiom (Luce, 1959); for each token categorised by each listener, the expected accuracy for that token is the model's posterior for the category selected by each listener. We took the average log posterior of all responses to get the average likelihood for the entire experiment under each model.

The first point that stands out from the visual comparisons is that models that incorporate perceptual noise fit the perceptual data better than those that do not. This itself indicates that perception of acoustic stimuli is not entirely faithful to the bottom-up signal but is inferred through a combination of what listeners actually perceived and their existing knowledge of the underlying linguistic category (Kronrod et al., 2016). For the univariate VOT models, the difference is most noticeable from the flatter slopes of the IOs indicating greater uncertainty in listener categorisations. The second pattern is that models trained with VOT and F0 cues (multiple cues) are better fits overall than models trained on a single cue. This trend is expected given the literature that report F0 reliably covarying with the voicing of stop consonants (House & Fairbanks, 1953; Ohde, 1984). When VOT fails to provide sufficient support to voicing status, F0 has been found to influence listeners' categorisation behaviour (Abramson & Lisker, 1985; Idemaru & Holt, 2011; Whalen, Abramson, Lisker, & Mody, 1993; Winn, Chatterjee, & Idsardi, 2013; **burchilljaeger2023?**). This further speaks to the advantage of multivariate ideal observers because they assess the likelihood of a cue observation under a given category relative to the joint distributions of all relevant cues.

```
## Warning in geom_errorbar(data = d.bootstrap %>% group_by(io.type, gender) %>% : Ignoring unl
```

(ref:comparing-likelihoods-of-perception-data-under-each-bivariate-IO)

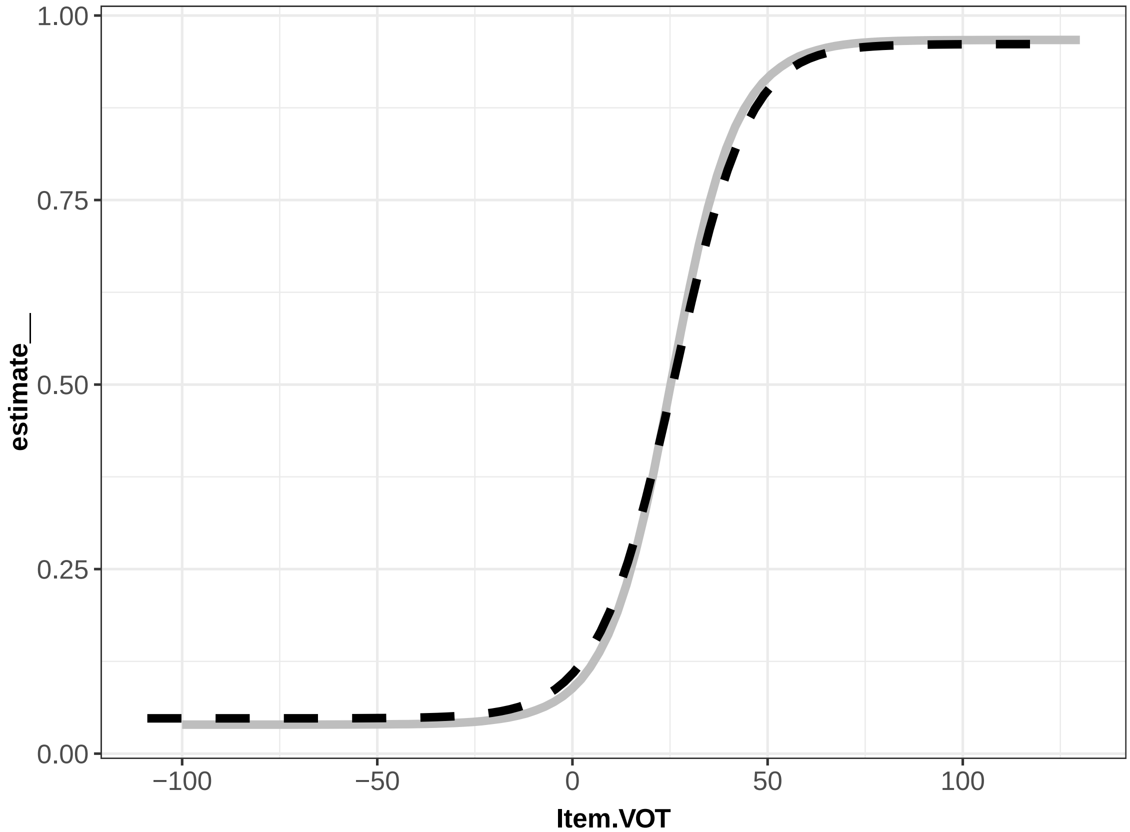

Figure 3. visualising difference between uncentered and centered exposure fit (dashed line).

3 EXPERIMENT 2: Listeners' adaptation to an unfamiliar talker

The aim of experiment 2 was to investigate the incremental changes in listener categorization when perceiving speech of an unfamiliar talker with cue-to-category mappings characterised by varying degrees of typicality of an L1-US English talker. Listeners performed a task similar to that of experiment 1, that is, they heard isolated words on a /d/ - /t/ continuum and were required to select the word they heard. Unlike experiment 1 where all listeners categorised stimuli on a single uninformative continuum, listeners in experiment 2 were divided into 3 groups with each group exposed to different VOT distributions that were informative of the talker's realisations of /d/ and /t/.

We approximated a "typical" talker through the combined parameters estimated from the perceptual responses in experiment 1 and a database of L1-US English /d/ and /t/ productions

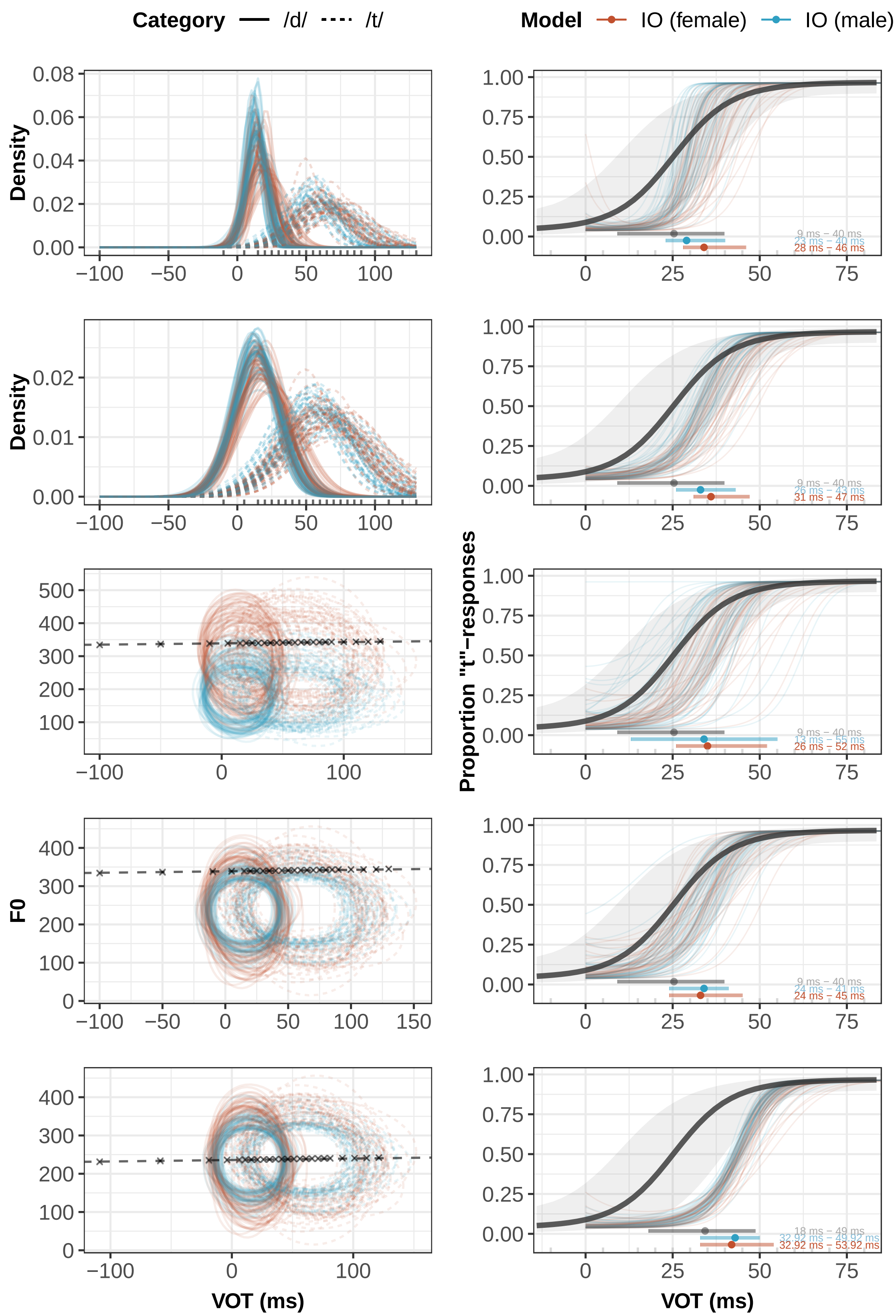

Figure 4. **Right column:** Comparing predicted vs. observed categorization functions for Experiment 1. The black line and interval show the psychometric fit and 95% CI for Experiment 1 marginalizing over all random effects. Each thin line shows the prediction of a single talker-specific ideal observer derived from a database of word-initial stop productions (data: Chodroff & Wilson, 2017; data preparation & model code: X. Xie, Jaeger, & Kurumada, 2022). The lapse rate and response bias for the ideal observers was set to match the MAP estimates of the psychometric model. For ease of comparisons, horizontal point ranges show the PSE and its 95% CI after discounting

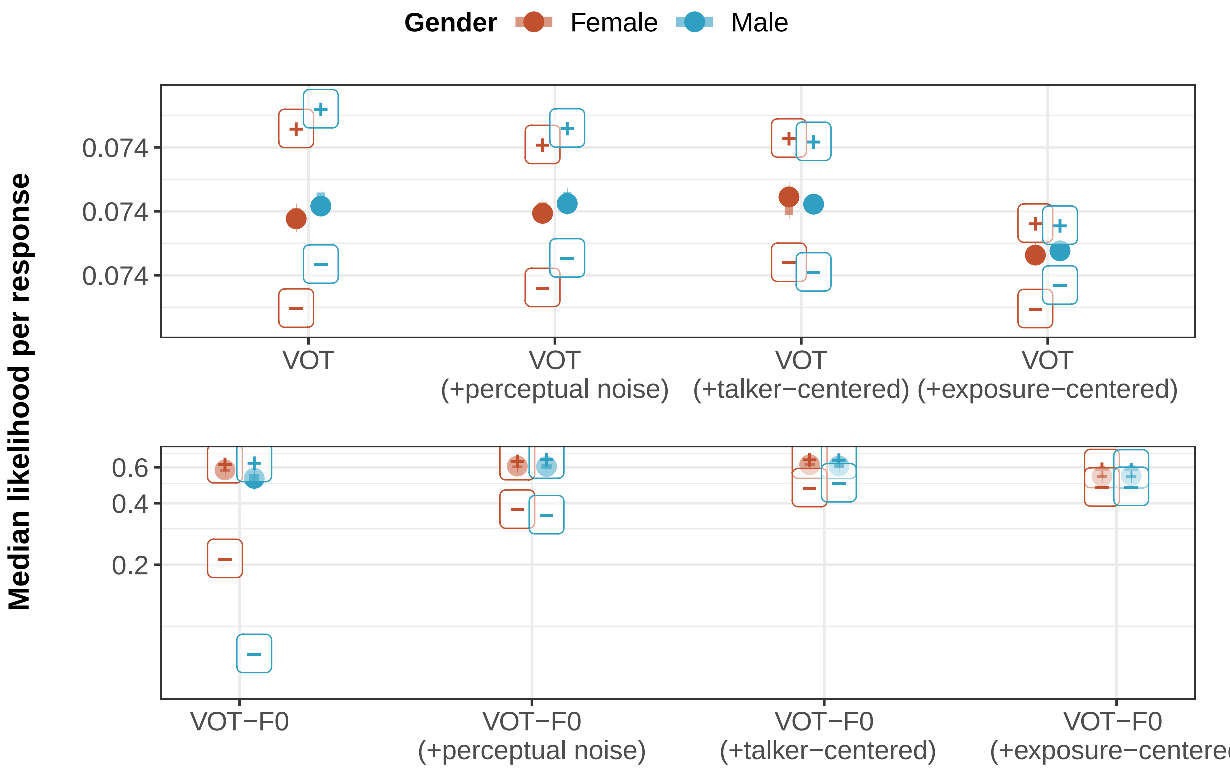

Figure 5. (ref:comparing-likelihoods-of-perception-data-under-each-bivariate-IO)

(**Xie?**). From this estimated baseline distribution (+0ms), we shifted the distribution by +10ms, and by +40ms, yielding three exposure talker conditions. To investigate the state of listener expectations as they move from having no information about how a new talker realises /d/s and /t/s to progressively more information about the talker's pronunciations we implement identical test blocks (i.e. test stimuli in identical locations) within and across conditions before, during, and after informative exposure. Under Bayesian ideal adaptor inferential processes, listeners' weighting of their prior beliefs about the category means and variances will determine the speed at which adaptation occurs. Motivated by prior work in supervised and unsupervised learning within lab contexts that repeatedly show adaptation to be a rapid process Kleinschmidt & Jaeger (2012) we made the decision to test our participants early on in the experiment.

Previous studies were not designed to investigate incremental adaptation in this manner as they lacked designated test blocks; listeners' categorisation functions were instead estimated over portions of the exposure trials which ignores the possibility that not all participants had been exposed to the full distributional information at the trial cut-off point (although that would have been the case by the end of the experiment). With our novel design we gain better resolution at every testing point, since each participant would have heard the same number of VOT items at the beginning of a given test block. The other advantage is that identical test blocks across conditions standardises the assessment of behavioural changes between groups yielding more accurate comparisons. We specifically included a pre-exposure test block with a similar aim to experiment 1 – in order to capture the implicit expectations of listeners about the cue-to-category mappings of US English /d/ and /t/. We later compare this block with the behavioural results of experiment 1.

Another notable innovation we bring to this study in conjunction with the use of qualitatively more human-sounding stimuli (as described in section 2.X), relates to the parameters of the exposure distributions. Prior studies of this type simulate the voiced-voiceless distributions by exposing listeners to category distributions that are symmetrical and equivalent between categories. It is however, unlikely that listeners encounter this in real life as evidenced from production data (**chodroffstructure?**). By generating distributions that are closer in form to that of real data we hope to improve the ecological validity of the results.

3.1 Methods

3.1.1 Participants

Participants were recruited over the Prolific platform and experiment data (but not participant profile data) were collected, stored, and via proliferate ((**schuster?**)). They were paid \$8.00 each (for a targeted remuneration of \$9.60/hour). The experiment was visible to participants following a selection of Prolific's available pre-screening criteria. Participants had to (1) have US nationality, (2) report to only know English, and (3) had not previously participated in any experiment from our lab on Prolific.

126 L1 US English listeners (male = 60, female = 59, NA = 3; mean age = 38 years; SD age = 12 years) completed the experiment. Due to data transfer errors 4 participants' data were not stored and therefore not included in this analysis. To be eligible, participants had to confirm that they (1) spent at least the first 10 years of their life in the US speaking only English, (2) were in a quiet place and free from distractions, and (3) wore in-ear or over-the-ears headphones

that cost at least $15.

3.1.2 Materials

A subset of the materials described in experiment 1 were used, in particular three continua of the minimal pairs, dill-till, din-tin, and dip-tip. The dim-tim continuum was omitted in order to keep the pairs as distinguishable as possible.

We employed a multi-block exposure-test design 6 which enabled the assessment of listener perception before informative exposure as well as incrementally at intervals during informative exposure (after every 48 exposure trials). To have a comparable test between blocks and across conditions, test blocks were made up of a uniform distribution of 12 VOT stimuli (-5, 5, 15, 25, 30, 35, 40, 45, 50, 55, 65, 70), identical across test blocks and between conditions. Each of the test tokens were presented once at random. The test blocks were kept short to avoid cancelling out any distributional learning effects after each exposure. After the final exposure block we tripled the number of test blocks to increase the statistical power to detect exposure induced changes.

The conditions were created by first obtaining the baseline distribution (+0ms shift) and then shifting that distribution by +10ms and by +40ms to create the remaining two conditions.

The +0ms shift condition was estimated from the fitted point of subjective equality (PSE) from experiment 1. The PSE corresponds to the VOT measurement that was perceived as the most ambiguous by participants in experiment 1 (i.e. the stimulus that elicited equal probability of being categorised as /d/ or /t/) thus marking the categorical boundary. The PSE is where the likelihoods of both categories intersect and have equal density (we assumed Gaussian distributions and equal prior probability for each category) [SOMETHING HERE ABOUT GAUSSIANS BEING A CONVENIENT ASSUMPTION?]. To limit the infinite combinations of likelihoods that could intersect at this value, we set the variances of the /d/ and /t/ categories based on parameter estimates (X. Xie et al. (2022)) obtained from the production database of Chodroff and Wilson (2017). To each variance value we added an 80ms noise variance following (Kronrod et al. (2016)) to account for variability due to perceptual noise since these likelihoods were estimated from perceptual data. We took an additional degree of freedom of setting the *distance between the means* of the categories at 46ms; this too was based on the population parameter estimates derived from analyses of the production database. The means of both categories were then obtained through a grid-search process to find the likelihood distributions that crossed at 25ms VOT (see XX of SI for details on this procedure).

The distributional make up was determined through a process of sampling tokens from a discrete normal distribution (available through the `extraDistr` package in R). [EXPLAIN WHAT DISCRETE NORMAL SAMPLING GIVES] discretised normal distributions are approximation…

For each exposure block 8 VOT tokens of each minimal pair item were sampled from discrete normal distributions of each category of the +0ms condition, giving 24 /d/ and 24 /t/ (48 critical trials) per block. Additionally, each exposure block contained 2 instances of 3 catch items, giving 6 catch trials per block. The sampled VOT tokens were increased by a margin of +10ms and +40 ms for the remaining two conditions. Three variants of each condition list were created so that exposure blocks followed a latin-square order.

Lastly, half of the exposure trials were randomly assigned as labelled trials. In labelled trials, participants receive clear information of the word's category as both orthographic options will always begin with the intended sound. For example if a trial was intended to be "dill" then the two image options will either be "dill" and "dip" or "dill" and "din". Test trials were always unlabelled.

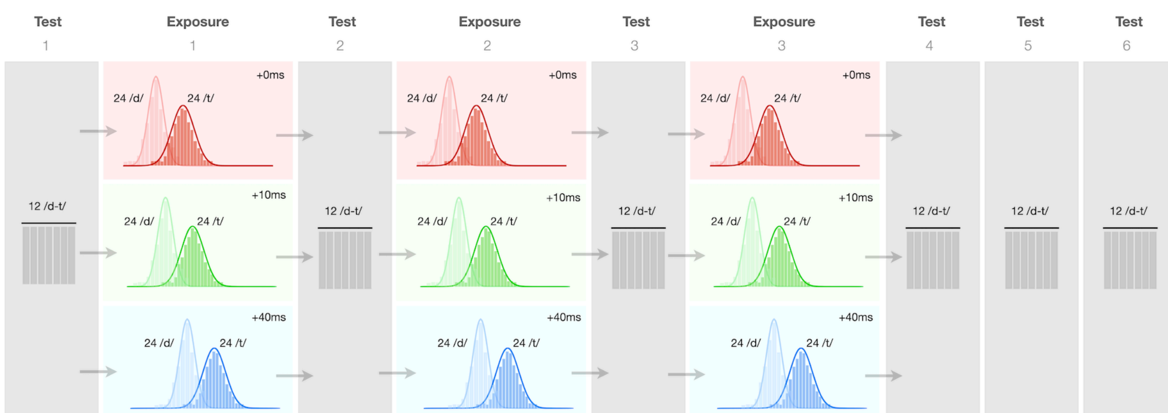

Figure 6. Experiment 2 multi-block design. Test blocks in grey comprised identical stimuli within and between conditions

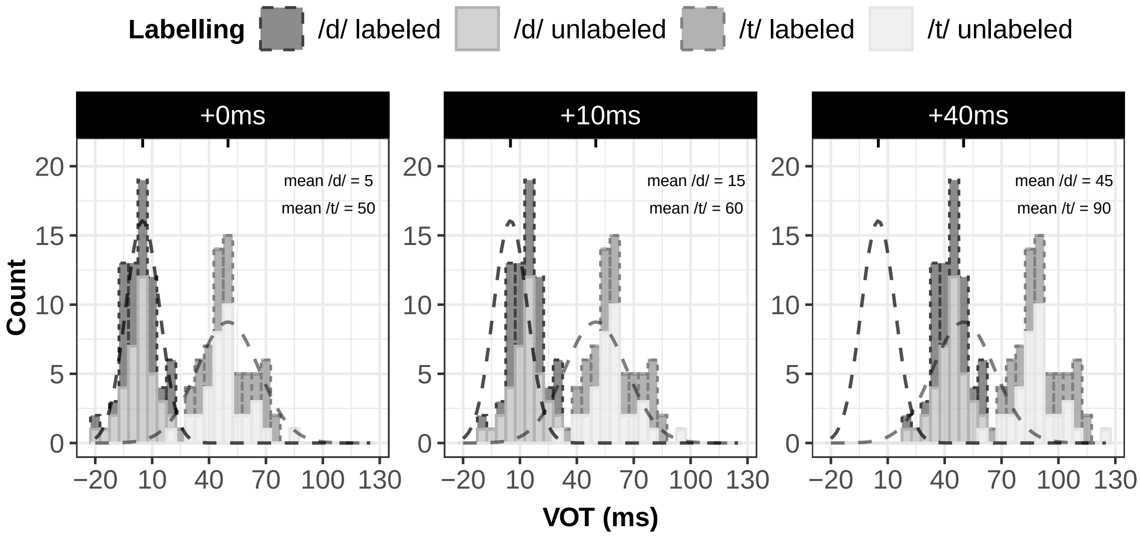

Figure 7

3.1.3 Procedure

The code for the experiment is available as part of the OSF repository for this article. A live version is available at (https://www.hlp.rochester.edu/FILLIN-FULL-URL). The first page of the experiment informed participants of their rights and the requirements for the experiment: that they had to be native listeners of English, wear headphones for the entire duration of the experiment, and be in a quiet room without distractions. Participants had to pass a headphone test, and were asked to keep the volume unchanged throughout the experiment. Participants could only advance to the start of the experiment by acknowledging each requirement and consenting to the guidelines of the Research Subjects Review Board of the University of Rochester.

On the next page, participants were informed about the task for the remainder of the experiment. They were informed that they would hear a female talker speak a single word on each trial, and had to select which word they heard. They were also informed that they needed to click a green button that would be displayed during each trial when it “lights up” in order to hear the recording of the speaker saying the word. Participants were instructed to listen carefully and answer as quickly and as accurately as possible. They were also alerted to the fact that the recordings were subtly different and therefore may sound repetitive. This was done to encourage their full attention.

The trials were presented in the same way as in experiment 1 except that the audio playback was controlled by the participant. This additional step was implemented to increase participant attention to the stimuli. The placement of the image presentations were counter-balanced across participants.

Participants underwent 234 trials which included 6 catch trials in each exposure block (18 in total). Participants were given the opportunity to take breaks after every 60 trials during exposure blocks. Participants took an average of 17 minutes (SD = 9) to complete the 234 trials, after which they answered a short survey about the experiment.

```
## # A tibble: 5 x 3
## # Groups:   Exclude_participant.due_to_VOT_slope [2]
##   Exclude_participant.due_to_VOT_slope Condition.Exposure `n()`
##   <lgl>                                <chr>              <int>
## 1 FALSE                                Shift0                41
## 2 FALSE                                Shift10               40
## 3 FALSE                                Shift40               39
## 4 TRUE                                 Shift0                 1
## 5 TRUE                                 Shift10                1
```

3.1.4 Exclusions

We excluded from analysis participants who committed more than 3 errors out of the 18 catch trials (<84% accuracy, N = 1), participants who committed more than 4 errors out of the 72 catch trials (<94% accuracy, N = 0), participants with an average reaction time (RT) more than three standard deviations from the mean of the by-participant means (N = 0), and participants who reported not to have used headphones (N = 0) or not to be native (L1) speakers of US English (N = 0).

In addition, participants' categorization during the early phase of the experiment were scrutinised for their slope orientation and their proportion of "t"-responses at the least ambiguous locations of the VOT continuum. The early phase of the experiment was defined as the first 36 trials and the least ambiguous locations were defined as -20ms from the empirical mean of the /d/ category and +20ms from the empirical mean of the /t/ category. These means were taken from the production data estimates by X. Xie et al. (2022). For the remaining participants, trials that were more than three SDs from the participant's mean RT were excluded from analysis (1.7%). Finally, we excluded participants (N = 0) who had less than 50% data remaining after these exclusions.

3.2 Behavioral results

We first present participants' categorisation responses. Given that this experiment was designed to give pre-exposure test data, we run an analysis on test block 1 that is similar to the IO analysis of experiment 1.

3.2.1 Analysis approach

3.3 Regression analysis

The regression analysis addresses two main questions: Do participants shift their categorisation behaviour in an incremental fashion, i.e. do they exhibit categorisation behaviour that draws closer to the ideal categorisation function with each successive exposure block? Are the differences in shifts between the conditions proportional to the magnitude of the shifts between exposure distributions i.e. is the PSE of the +40ms condition 3 times that of the +10ms condition?

As with experiment 1 we fit a Bayesian mixed-effects psychometric model with lapse and perceptual components. Continuous predictors were standardised to twice the standard deviation and priors and sampling parameters were identical to those specified in experiment 1.

To analyse the incremental effects of exposure condition on the proportion of /t/ responses at test, the perceptual model contained exposure condition (backward difference coded, comparing the +10ms against the +0ms shift 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. For the perceptual model, "t"-responses were regressed on the three-way interaction of VOT, condition, and block. Random effects were modelled with varying intercepts and slopes by participant and varying intercepts and slopes by minimal pair item. The lapsing model which estimates participant bias on trials with attention lapses was fitted without an intercept but with an offset [how does one describe this? what does offset(0)

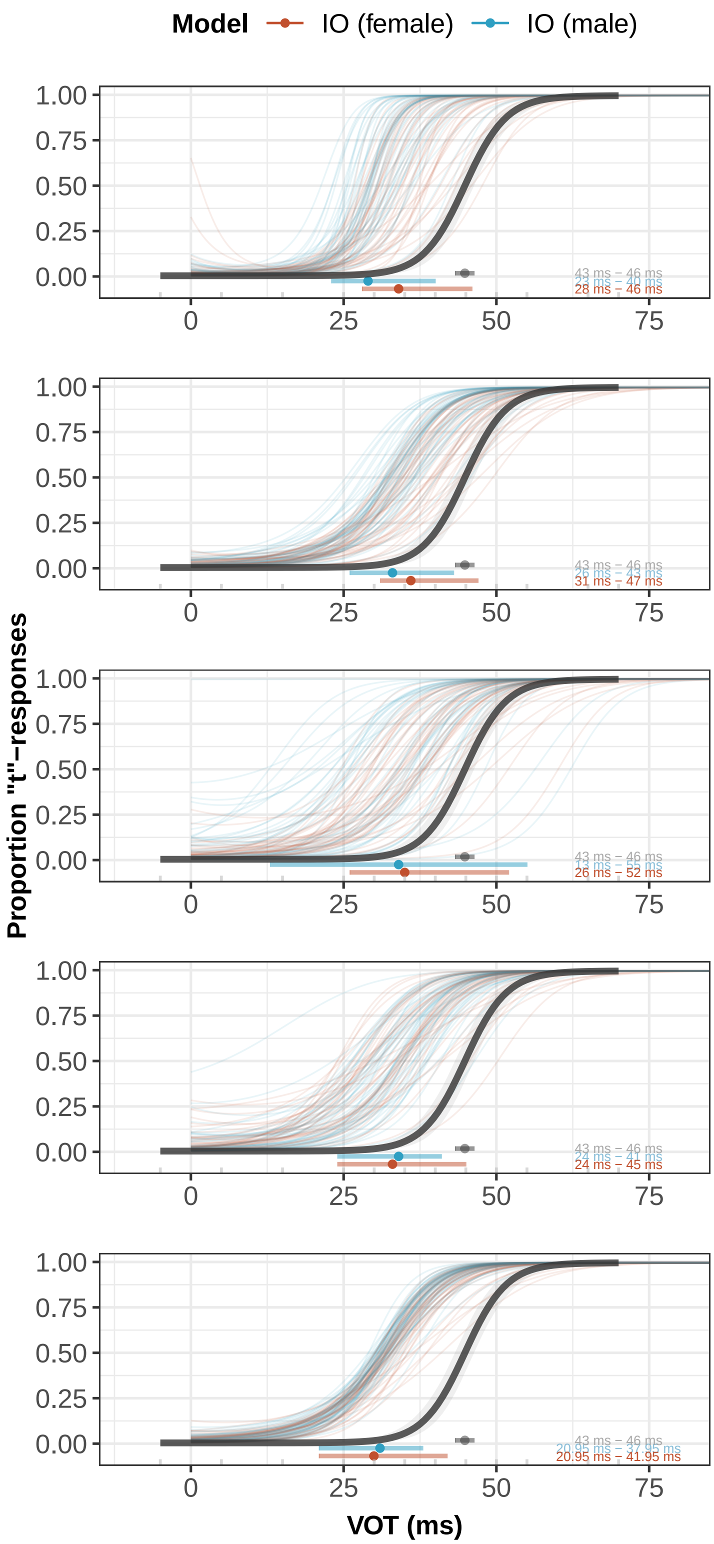
Model
IO (female)
IO (male)
Proportion "t"–responses
VOT (ms)
1.00
0.75
0.50
0.25
0.00
0
25
50
75
43 ms – 46 ms
28 ms – 46 ms
31 ms – 47 ms
13 ms – 55 ms
26 ms – 52 ms
24 ms – 45 ms
20.95 ms – 41.95 ms

Figure 8

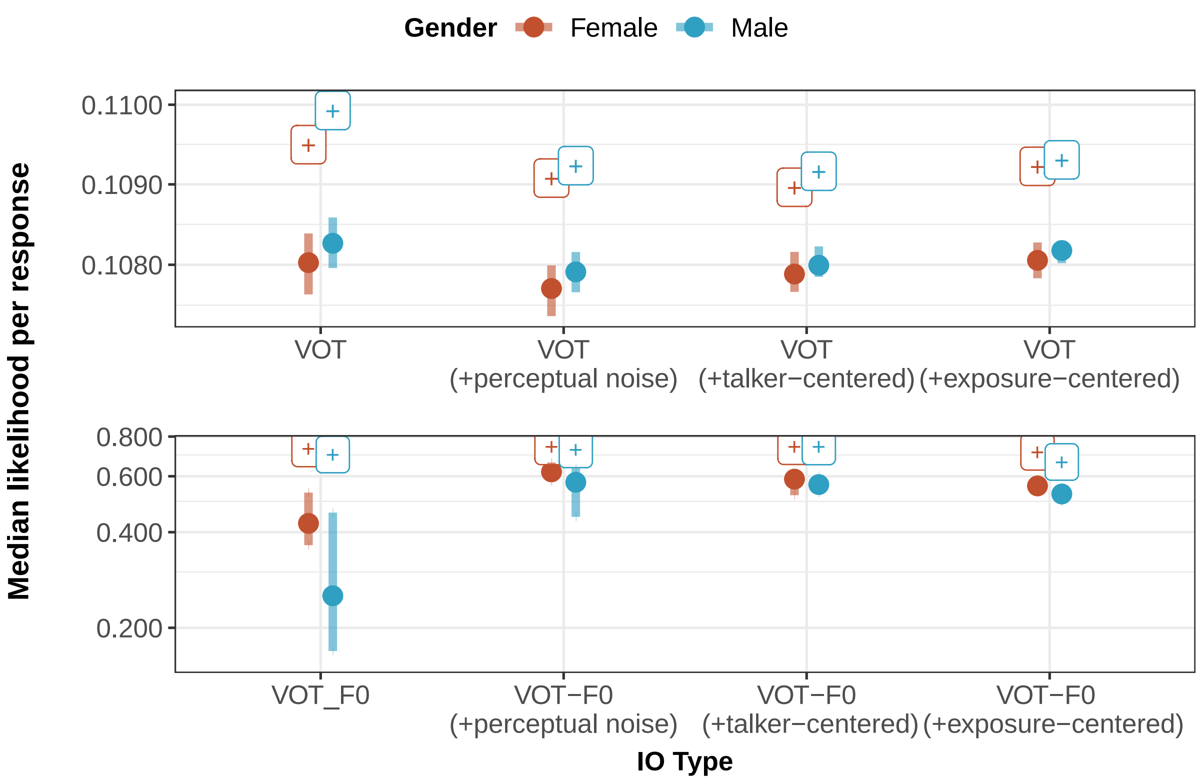

Figure 9

represent]. Finally, a population-level intercept was fitted to estimate the lapse rate. Random effects for the lapsing model and lapse rates were not fitted to limit the number of parameters and to ensure model convergence.

3.3.1 Expectations

Given previous findings of Kleinschmidt and Jaeger (2016) we expected participants in the various exposure conditions to shift their average categorization functions towards the direction of the ideal categorization function implied by their respective exposure distributions. We expected the differences between the groups to be most pronounced after the final exposure block as they would have had the complete exposure to all the tokens that make up the exposure distributions. This follows from predictions of incremental Bayesian belief-updating – that listeners would integrate their prior expectations with the current input to infer the present talker's cue-to-category-mapping (the posterior distribution). Also based on previous findings, we expected the +40ms group to not fully converge on the ideal categorization function as it was previously found that the further an exposure talker's cue distributions deviated from a *typical* talker's, the further the distance of categorization function from the ideal boundary. We therefore expected to see differences in categorizations between the +10ms and +40ms conditions such that listeners in the +40ms condition would shift more than those in the +10ms condition but to have an average categorization function located to the left of the ideal function. (Kleinschmidt & Jaeger, 2016).

Fig. XX summarizes participants' categorization functions across the different test blocks. A first point to note are the average categorization functions of the respective conditions before exposure to the talker. As depicted in the first panel, the average functions converge on the same boundary or PSE (4xms, CI =) which suggests that participants largely had similar expectations about the cue distribution corresponding to /d/ and /t/ for this type of talker. What it also shows is that in setting our baseline condition we may have underestimated the perceived boundary for our test stimuli by approximately 20ms which implies that the +10ms shift and the +40ms shift were in fact -10ms and +20ms respectively.[ELABORATION]

There was a main effect of VOT $\hat{\beta} = 16.9$ 95%-CI: 13.6 to 20.5; Bayes factor: Inf 90%-CI : 14.29 to 19.71; participants were more likely to respond "t" as VOT increased. Condition had a main effect on responses such that with larger shifts, participants on average responded with fewer "t"s. Additionally, the difference in average "t" responses between the +40ms and +10ms conditions ($\hat{\beta} =$ -2.4 95%-CI: -4 to -0.9; Bayes factor: 185.05 90%-CI : -3.67 to -1.2 reduction in log-odds) was *larger* than the difference between the +10 and +0 conditions ($\hat{\beta} =$ -1 95%-CI: -2.5 to 0.5; Bayes factor: 13.79 90%-CI : -2.21 to 0.15 reduction in log-odds). Qualitatively, the results indicate listeners adjust their expectations to align with the statistics of the exposure talker, consonant with previous findings of studies employing this paradigm (e.g., Clayards et al. (2008b); Kleinschmidt and Jaeger (2016); Theodore and Monto (2019)).

While there was weak evidence for a main effect of block its interaction with condition revealed how participants in the respective exposure groups responded as they progressively received more informative input. Most of the change took place after the first exposure block. Participants in the +10ms condition responded with fewer "ts" compared to participants in the +0ms condition in test block 2 relative to that in test block 1 ($\hat{\beta} =$ -1.4 95%-CI: -3.8 to 1.1; Bayes factor: 8.78 90%-CI : -3.32 to 0.54). The difference between the +40ms and +10ms condition in test block 2 relative to that in block 1 was more pronounced, reflecting the wider separation between the two exposure conditions in block 2 ($\hat{\beta}$ = -2.2 95%-CI: -4.9 to 0.4; Bayes factor: 23.62 90%-CI : -4.32 to -0.15).

In test block 3, the difference in average log-odds between conditions +0ms and +10ms, relative to block 2 was *positive* such that the difference between the two conditions in test block 3 was smaller than the corresponding difference in block 2 ($\hat{\beta}$ = 1 95%-CI: -1.6 to 3.9; Bayes factor: 4.77 90%-CI : -1.02 to 3.12). In test blocks 4 and 5, the average log-odds difference between +0ms and +10ms increased marginally when compared to the preceding block, respectively (as indicated by the negative signs of the estimates; see table xx) while in test block 6 the difference between the two exposure conditions narrowed substantially. Looking at the the difference between the +40ms and +10ms conditions, this continued to widen in test blocks 3 and block 4 relative to their respective preceding blocks, albeit by progressively smaller increments. This widening trend would then reverse in test blocks 5 and 6. In all, the respective conditions hit their maximal shifts between blocks 2 and 3 and began to display a reversal of the exposure effects by the end of block 4. This “unlearning” of the exposure distribution, observed in the final 3 test blocks was expected given previous findings that distributional learning effects can begin to dissipate with prolonged testing with tokens from a uniform distribution.

An examination of the block-by-block changes in the intercepts and slopes of the respective conditions, confirmed that the changes in categorization behaviour were driven predominantly by changes in the intercept (fig xx). the slopes of all 3 conditions in test block 4, which immediately follows the final exposure block, and where participants would have had full exposure to their respective distributions, did not differ substantially from each other nor from their estimated starting point in test block 1. Conversely, the intercepts at these points in the experiment were more distinct from each other and from where they were estimated to be at test block 1.

In summary, the analysis shows that the groups diverged in their categorisation behaviour very early on in the experiment – only after 24 exposures to each category. This suggests a readiness to adapt to a new talker by integrating current input with prior expectations. This prompt shift was however tempered by participants reaching the limits of their adaptation almost as quickly; the +40ms condition for example achieved more than 95% of its maximal shift during the experiment in test block 2. Only a marginal change in categorization behaviour was observed after the second exposure block while the third exposure block barely resulted in further shifts. Glaringly, all three conditions undershot the ideal categorization boundaries implied by their respective exposure distributions: 14.5ms in the +0ms, 7.2ms in the +10ms, and 14.5ms in the +40ms conditions. Like this study's antecedent, the various exposure groups did

Under the Bayesian ideal adapter framework quick adaptation is characterised as listeners having weak beliefs in their prior cue means and variances. Listeners' strength in prior beliefs influences the speed of adaptation, and this is what we observed from the analysis so far. On the other hand, weak prior beliefs also predict that it would take few trials for listeners to converge on the implied categorisation boundary. But this is not what we observed in our data. Instead, listeners were held back and stayed close to their … mean As listeners adapt to new talkers either by shifting their expectations of the mean, by expanding the variance or by both, the strength of pri. We dig deeper into the behaviour of the participants by running IA analyses in the next section

```
## Warning: Removed 3 rows containing missing values (`geom_line()`).

## Warning: Removed 6 rows containing missing values (`geom_pointrange()`).
```

```
## # A tibble: 18 x 5
## # Groups:   Condition [3]
##    Condition Block lower median upper
##    <chr>     <chr> <dbl>  <dbl> <dbl>
##  1 0         1      35.6   44.0  53.0
##  2 0         3      32.4   39.8  47.8
##  3 0         5      33.7   40.1  46.8
##  4 0         7      30.9   39.5  48.6
##  5 0         8      33.5   40.5  48.8
##  6 0         9      36.8   41.2  46.3
##  7 10        1      36.5   46.0  56.2
##  8 10        3      37.8   43.6  50.2
##  9 10        5      35.8   42.8  50.5
```

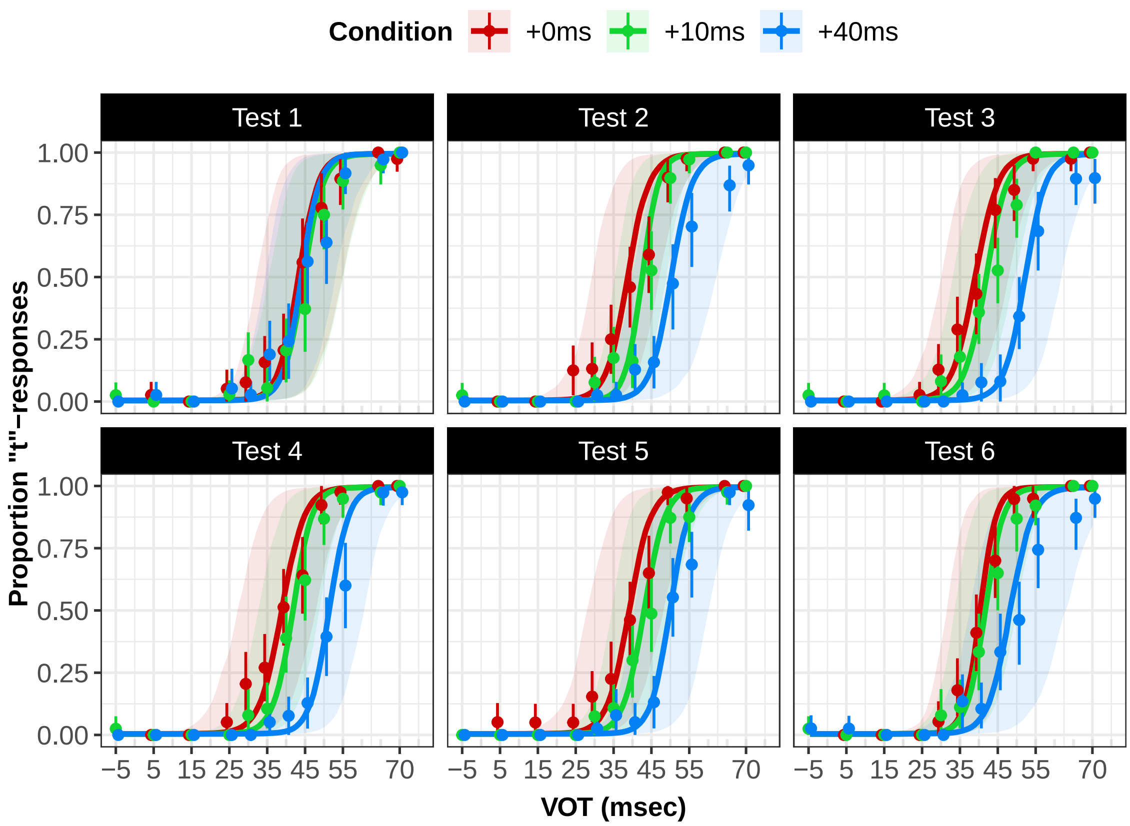

Figure 10

```
## 10 10        7       37.5    42.2   47.5
## 11 10        8       38.9    43.9   49.4
## 12 10        9       38.6    42.5   46.8
## 13 40        1       37.4    44.5   52.1
## 14 40        3       45.7    51.6   58.7
## 15 40        5       47.0    53.5   62.0
## 16 40        7       47.0    52.5   58.7
## 17 40        8       44.8    50.9   57.7
## 18 40        9       44.0    49.8   57.5
## Warning in tidy.brmsfit(fit_mix_nested, effects = "fixed"): some parameter names contain unc
## # A tibble: 40 x 2
##    ParticipantID `n()`
##            <dbl> <int>
```

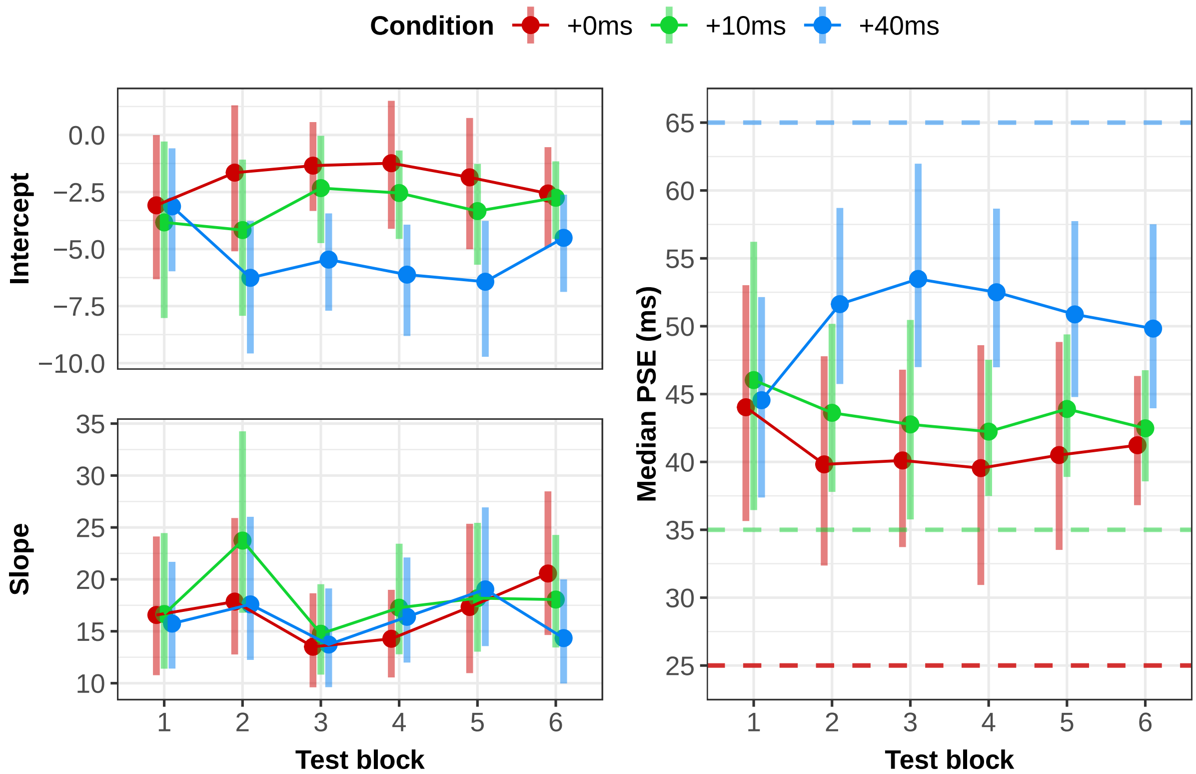

Figure 11

```
## 1               118    211
## 2               119    213
## 3               125    213
## 4               128    214
## 5               146    212
## 6               147    214
## 7               148    214
## 8               155    209
## 9               156    211
## 10              160    212
## # ... with 30 more rows
```

All data and code for this article can be downloaded fromhttps://osf.io/q7gjp/. This article is written in R markdown, allowing readers to replicate our analyses with the press of a button using freely available software (R, R Core Team, 2021a; RStudio Team, 2020), while changing any of the parameters of our models. Readers can revisit any of the assumptions we make—for example, by substituting alternative models of linguistic representations. The supplementary information (SI, §1) lists the software/libraries required to compile this document. Beyond our immediate goals here, we hope that this can be helpful to researchers who are interested in developing more informative experimental designs, and to facilitate the interpretation of existing results (see also Tan, Xie, & Jaeger, 2021).

4 General discussion

4.1 Methodological advances that can move the field forward

An example of a subsection.

5 References

Abramson, A. S., & Lisker, L. (1973). Voice-timing perception in spanish word-initial stops. *Journal of Phonetics*, *1*(1), 1–8.

Abramson, A. S., & Lisker, L. (1985). Relative power of cues: F0 shift versus voice timing. *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, 25–33.

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.

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Bache, S. M., & Wickham, H. (2020). *Magrittr: A forward-pipe operator for r.* Retrieved from https://CRAN.R-project.org/package=magrittr

Barth, M. (2022). *tinylabels: Lightweight variable labels.* Retrieved from https://cran.r-project.org/package=tinylabels

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

Bates, D., & Maechler, M. (2021). *Matrix: Sparse and dense matrix classes and methods.* Retrieved from https://CRAN.R-project.org/package=Matrix

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

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

Chang, W. (2022). *Webshot: Take screenshots of web pages*. Retrieved from https://CRAN.R-project.org/package=webshot

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*. https://doi.org/10.1515/lingvan-2017-0047

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., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008b). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804–809. https://doi.org/10.1016/j.cognition.2008.04.004

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008a). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809. https://doi.org/https://doi.org/10.1016/j.cognition.2008.04.004

Cole, J., Linebaugh, G., Munson, C., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, *38*, 167–184. https://doi.org/10.1016/j.wocn.2009.08.004

Csárdi, G., & Chang, W. (2021). *Processx: Execute and control system processes*. Retrieved from https://CRAN.R-project.org/package=processx

Daróczi, G., & Tsegelskyi, R. (2022). *Pander: An r 'pandoc' writer*. Retrieved from https://CRAN.R-project.org/package=pander

Dowle, M., & Srinivasan, A. (2021). *Data.table: Extension of 'data.frame'*. Retrieved from https://CRAN.R-project.org/package=data.table

Eddelbuettel, D., & Balamuta, J. J. (2018). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *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

Frick, H., Chow, F., Kuhn, M., Mahoney, M., Silge, J., & Wickham, H. (2022). *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. *The Annals of Applied Statistics*, *2*(4), 1360–1383.

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

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/

Hay, J., & Drager, K. (2010). *Stuffed toys and speech perception.*

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools.* Retrieved from https://CRAN.R-project.org/package=purrr

Henry, L., & Wickham, H. (2021). *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. (2020). *Ggstance: Horizontal 'ggplot2' components.* Retrieved from https://CRAN.R-project.org/package=ggstance

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

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America*, *25*(1), 105–113.

Hugh-Jones, D. (2021). *Latexdiffr: Diff 'rmarkdown' 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.

Jaeger, T. F. (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 NormalizationÖ an exemplar model. *Talker Variability in Speech Processing*, 145–165.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384.

Kassambara, A. (2020). *Ggpubr: 'ggplot2' based publication ready plots*. Retrieved from https://CRAN.R-project.org/package=ggpubr

Kay, M. (2022a). *ggdist: Visualizations of distributions and uncertainty*. https://doi.org/10.5281/zenodo.3879620

Kay, M. (2022b). *tidybayes: Tidy data and geoms for Bayesian models*. https://doi.org/10.5281/zenodo.1308151

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. https://doi.org/https://doi.org/10.1037/a0038695

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

Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, *23*(6), 1681–1712. https://doi.org/https://doi.org/10.3758/s13423-016-1049-y

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

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

Liu, L., & Jaeger, T. F. (2018a). Inferring causes during speech perception. *Cognition*, *174*, 55–70. https://doi.org/10.1016/j.cognition.2018.01.003

Liu, L., & Jaeger, T. F. (2018b). 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*, 1562–1588. https://doi.org/10.1037/xhp0000693

Luce, R. D. (1959). Individual choice behavior. In *Individual choice behavior.* (pp. 153, xii, 153–xii). John Wiley.

Maechler, M. (2021). *Diptest: Hartigan's dip test statistic for unimodality - corrected*. Retrieved from https://CRAN.R-project.org/package=diptest

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.

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

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.

Müller, K., & Wickham, H. (2021). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Nearey, T. M., & Hogan, J. T. (1986). Phonological contrast in experimental phonetics: Relating distributions of production data to perceptual categorization curves. *Experimental Phonology*, 141–161.

Neuwirth, E. (2022). *RColorBrewer: ColorBrewer palettes*. Retrieved from https://CRAN.R-project.org/package=RColorBrewer

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

Ohde, R. N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *The Journal of the Acoustical Society of America*, *75*(1), 224–230.

Ooms, J. (2021). *Magick: Advanced graphics and image-processing in r.* Retrieved from https://CRAN.R-project.org/package=magick

Ooms, J. (2022). *Curl: A modern and flexible web client for r.* Retrieved from https://CRAN.R-project.org/package=curl

Pedersen, T. L. (2022a). *Ggforce: Accelerating 'ggplot2'.* Retrieved from https://CRAN.R-project.org/package=ggforce

Pedersen, T. L. (2022b). *Patchwork: The composer of plots.* Retrieved from https://CRAN.R-project.org/package=patchwork

Pedersen, T. L., & Robinson, D. (2020). *Gganimate: A grammar of animated graphics.* Retrieved from https://CRAN.R-project.org/package=gganimate

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *In Frequency and the Emergence of Linguistic Structure* (pp. 137–157). John Benjamins.

R Core Team. (2021a). *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. (2021b). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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

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

Sievert, C. (2020). *Interactive web-based data visualization with r, plotly, and shiny.* Chapman; Hall/CRC. Retrieved from https://plotly-r.com

Slowikowski, K. (2021). *Ggrepel: Automatically position non-overlapping text labels with 'ggplot2'.* Retrieved from https://CRAN.R-project.org/package=ggrepel

Statisticat, & LLC. (2021). *LaplacesDemon: Complete environment for bayesian inference.* Bayesian-Inference.com. Retrieved from https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software

Stevens, S. S., & Volkmann, J. (1940). The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, *53*(3), 329–353.

Tan, M., Xie, X., & Jaeger, T. F. (2021). Using rational models to understand experiments on accent adaptation. *Frontiers in Psychology*, *12*, 1–19. https://doi.org/10.3389/fpsyg.2021.676271

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

Theodore, R. M., & Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions. *Psychonomic Bulletin & Review*, *26*(3), 985–992. https://doi.org/https://doi.org/10.3758/s13423-018-1551-5

Toscano, J. C., & McMurray, B. (2015). The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments. *Language, Cognition and Neuroscience*, *30*(5), 529–543.

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*.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth). New York: Springer. Retrieved from https://www.stats.ox.ac.uk/pub/MASS4/

Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1993). F 0 gives voicing information even with unambiguous voice onset times. *The Journal of the Acoustical Society of America*, *93*(4), 2152–2159.

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

Wickham, H. (2019b). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H. (2020). *Modelr: Modelling functions that work with the pipe*. Retrieved from https://CRAN.R-project.org/package=modelr

Wickham, H. (2021a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2021b). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

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. (2021). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., Hester, J., & Bryan, J. (2021). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

Wickham, H., & Seidel, D. (2022). *Scales: Scale functions for visualization*. Retrieved from https://CRAN.R-project.org/package=scales

Wilke, C. O. (2020). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=cowplot

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.

Winn, M. B., Chatterjee, M., & Idsardi, W. J. (2013). Roles of voice onset time and F0 in stop consonant voicing perception: Effects of masking noise and low-pass filtering. *Journal of Speech, Language, and Hearing Research : JSLHR*, *56*, 1097–1107. https://doi.org/10.1044/1092-4388(2012/12-0086)

Xie, X., Jaeger, T. F., & Kurumada, C. (2022). *What we do (not) know about the mechanisms underlying adaptive speech perception: A computational review*. https://doi.org/10.17605/OSF.IO/Q7GJP

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*.

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.

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.org/knitr/

Xie, Y. (2021). *Knitr: A general-purpose package for dynamic report generation in r*. Retrieved from https://yihui.org/knitr/

Xie, Y., & Allaire, J. (2022). *Tufte: Tufte's styles for r markdown documents*. Retrieved from https://CRAN.R-project.org/package=tufte

Zhu, H. (2021). *kableExtra: Construct complex table with 'kable' and pipe syntax*. Retrieved from https://CRAN.R-project.org/package=kableExtra

Supplementary information

Both the main text and these supplementary information (SI) are derived from the same R markdown document available via OSF. It is best viewed using Acrobat Reader. Some links and animations might not work in other PDF viewers.

§1 Required software

The document was compiled using `knitr` (Y. Xie, 2021) in RStudio with R:

```
##                _
## platform       x86_64-apple-darwin17.0
## arch           x86_64
## os             darwin17.0
## system         x86_64, darwin17.0
## status
## major          4
## minor          1.3
## year           2022
## month          03
## day            10
## svn rev        81868
## language       R
## version.string R version 4.1.3 (2022-03-10)
## nickname       One Push-Up
```

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 4.1.3; R Core Team, 2021b) and the R-packages *}broom* [@}R-broom], *assertthat* (Version 0.2.1; Wickham, 2019a), *brms* (Version 2.18.0; Bürkner, 2017, 2018, 2021), *broom.mixed* (Version 0.2.9.4; Bolker & Robinson, 2022), *cowplot* (Version 1.1.1; Wilke, 2020), *curl* (Version 4.3.3; Ooms, 2022), *data.table*

(Version 1.14.8; Dowle & Srinivasan, 2021), *diptest* (Version 0.76.0; Maechler, 2021), *dplyr* (Version 1.1.0; Wickham, François, Henry, & Müller, 2021), *forcats* (Version 1.0.0; Wickham, 2021a), *gganimate* (Version 1.0.8; Pedersen & Robinson, 2020), *ggdist* (Version 3.2.1; Kay, 2022a), *ggforce* (Version 0.4.1; Pedersen, 2022a), *ggplot2* (Version 3.4.1; Wickham, 2016), *ggpubr* (Version 0.5.0; Kassambara, 2020), *ggrepel* (Version 0.9.2; Slowikowski, 2021), *ggstance* (Version 0.3.6; Henry, Wickham, & Chang, 2020), *kableExtra* (Version 1.3.4; Zhu, 2021), *knitr* (Version 1.42; Y. Xie, 2015), *LaplacesDemon* (Version 16.1.6; Statisticat & LLC., 2021), *latexdiffr* (Version 0.1.0; Hugh-Jones, 2021), *linguisticsdown* (Version 1.2.0; Liao, 2019), *lme4* (Version 1.1.31; Bates, Mächler, Bolker, & Walker, 2015), *lmerTest* (Version 3.1.3; Kuznetsova, Brockhoff, & Christensen, 2017), *lubridate* (Version 1.9.0; Grolemund & Wickham, 2011), *magick* (Version 2.7.3; Ooms, 2021), *magrittr* (Version 2.0.3; Bache & Wickham, 2020), *MASS* (Version 7.3.58.2; Venables & Ripley, 2002), *Matrix* (Version 1.5.1; Bates & Maechler, 2021), *modelr* (Version 0.1.10; Wickham, 2020), *pander* (Version 0.6.5; Daróczi & Tsegelskyi, 2022), *papaja* (Version 0.1.1.9,001; Aust & Barth, 2020), *patchwork* (Version 1.1.2; Pedersen, 2022b), *phonR* (Version 1.0.7; McCloy, 2016), *plotly* (Version 4.10.1; Sievert, 2020), *posterior* (Version 1.4.0; Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2021), *processx* (Version 3.8.0; Csárdi & Chang, 2021), *purrr* (Version 1.0.1; Henry & Wickham, 2020), *RColorBrewer* (Version 1.1.3; Neuwirth, 2022), *Rcpp* (Eddelbuettel & Balamuta, 2018; Version 1.0.10; Eddelbuettel & François, 2011), *readr* (Version 2.1.3; Wickham, Hester, & Bryan, 2021), *rlang* (Version 1.1.0; Henry & Wickham, 2021), *rsample* (Version 1.1.1; Frick et al., 2022), *scales* (Version 1.2.1; Wickham & Seidel, 2022), *stringr* (Version 1.5.0; Wickham, 2019b), *tibble* (Version 3.2.1; Müller & Wickham, 2021), *tidybayes* (Version 3.0.3; Kay, 2022b), *tidyr* (Version 1.3.0; Wickham, 2021b), *tidyverse* (Version 1.3.2; Wickham et al., 2019), *tinylabels* (Version 0.2.3; Barth, 2022), *tufte* (Version 0.12; Y. Xie & Allaire, 2022), and *webshot* (Version 0.5.4; Chang, 2022). 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.

§2 Overview

§2.1 Overview of data organisation

§3 Stimuli generation for perception experiments

§3.1 Recording of audio stimuli

§3.2 Annotation of audio stimuli

§3.3 Synthesis of audio stimuli

- acoustic plots

§4 Web-based experiment design procedure

§4.1 Experiment 1

§4.1.1 Making exposure conditions

§4.1.2 Exclusions analysis

§4.1.3 Regression analysis - model selection

```
## Warning in geom_line(data = fit_mix_f0_data %>% group_by(sVOT) %>% summarise(estimate__ = me
```

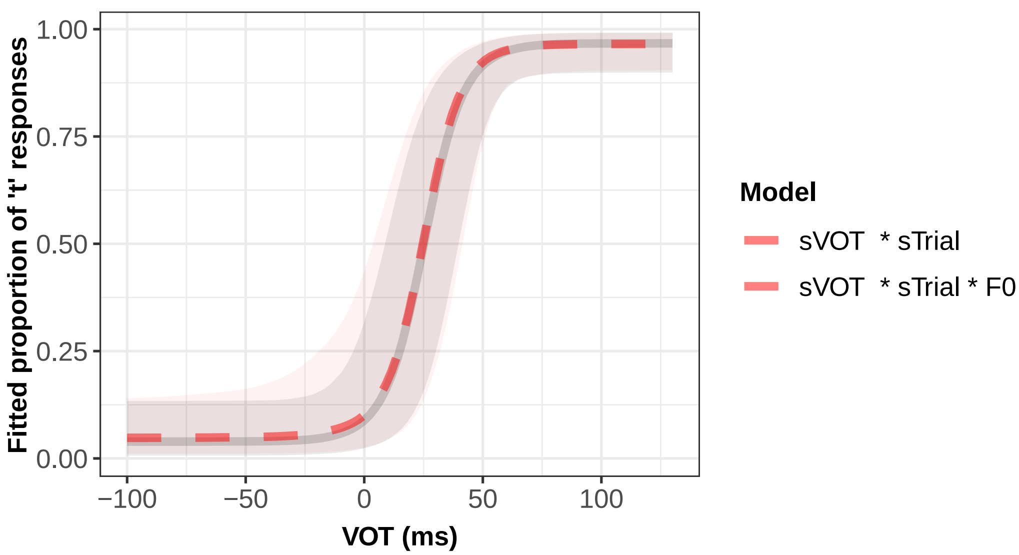

Figure 12. Expected effect of VOT interacting with trial on categorisation from model: 1 + (sVOT + sFO) * sTrial shown as red dashed line with pink shaded CI. Grey line and shaded area represents effects of VOT interacting with trial from model: 1 + sVOT * sTrial

§4.2 Experiment 2

§4.2.1 Making exposure conditions

§4.2.2 Exclusions analysis

- reaction time plots
- catch trial performance plots ### Regression analysis - model selection

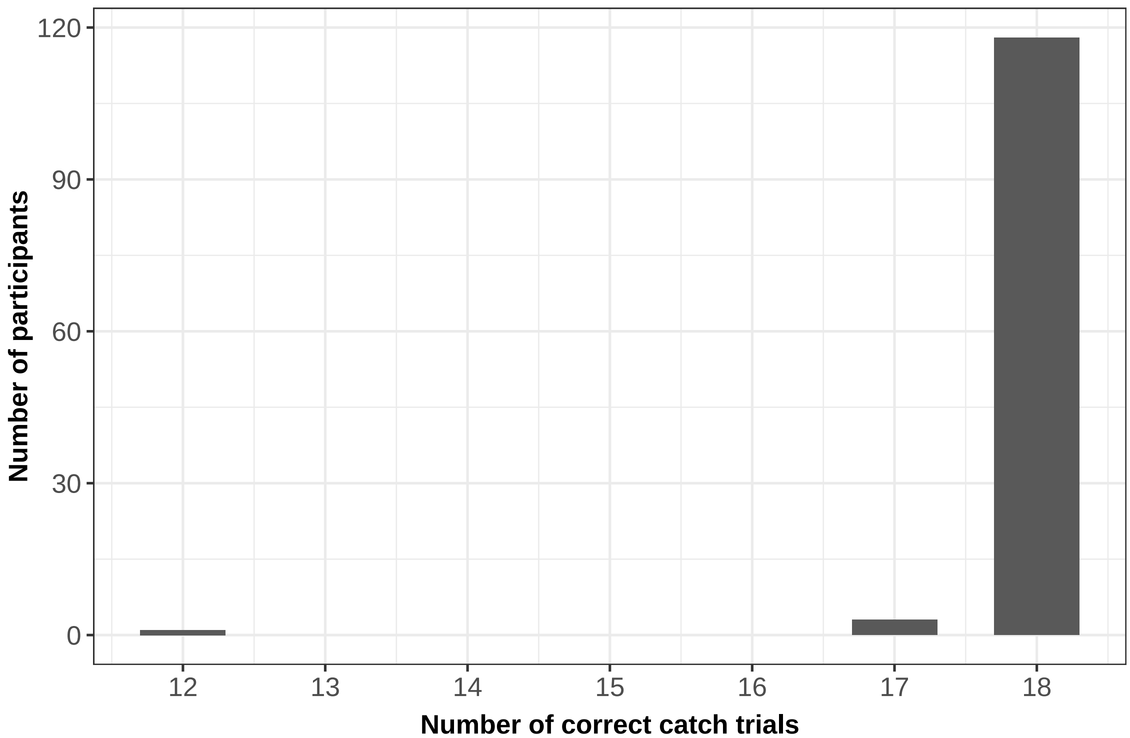

Figure 13

-labelled trial performance plots

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in dply
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()` always return
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

§4.3 Ideal observer training

We train the IOs on cue distributions extracted from an annotated database of XX L1 US-English talkers' productions (Chodroff and Wilson (2017)) of word initial stops. We apply Bayes' theorem to derive the IOs' posterior probability of categorising the test stimuli as "t". This is defined as

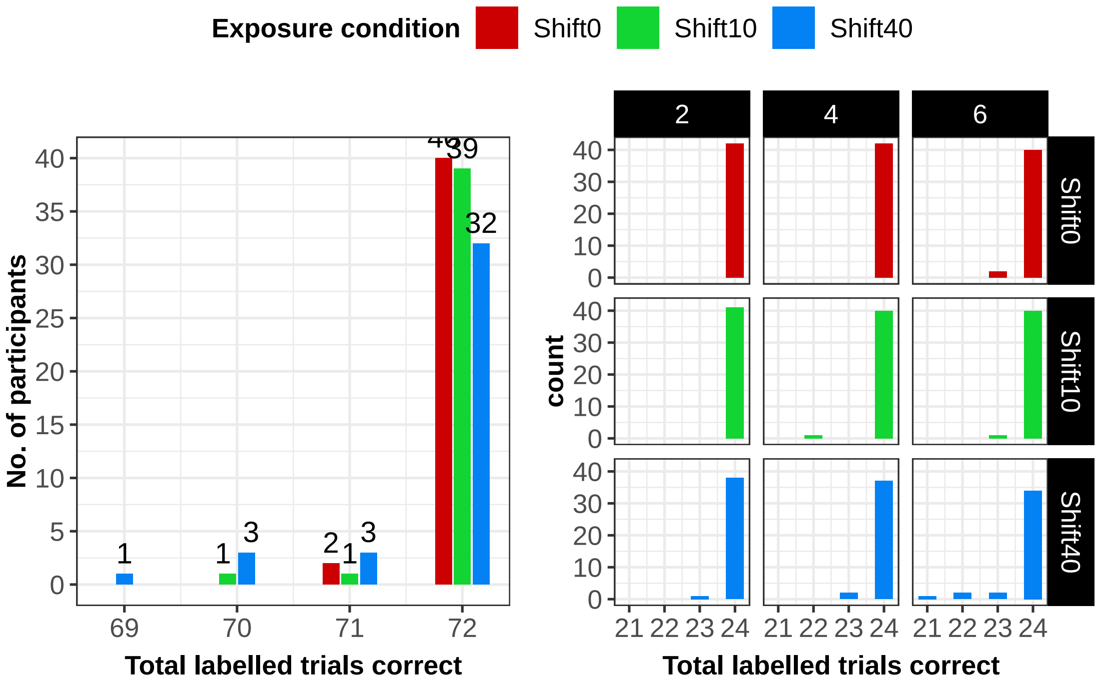

Figure 14

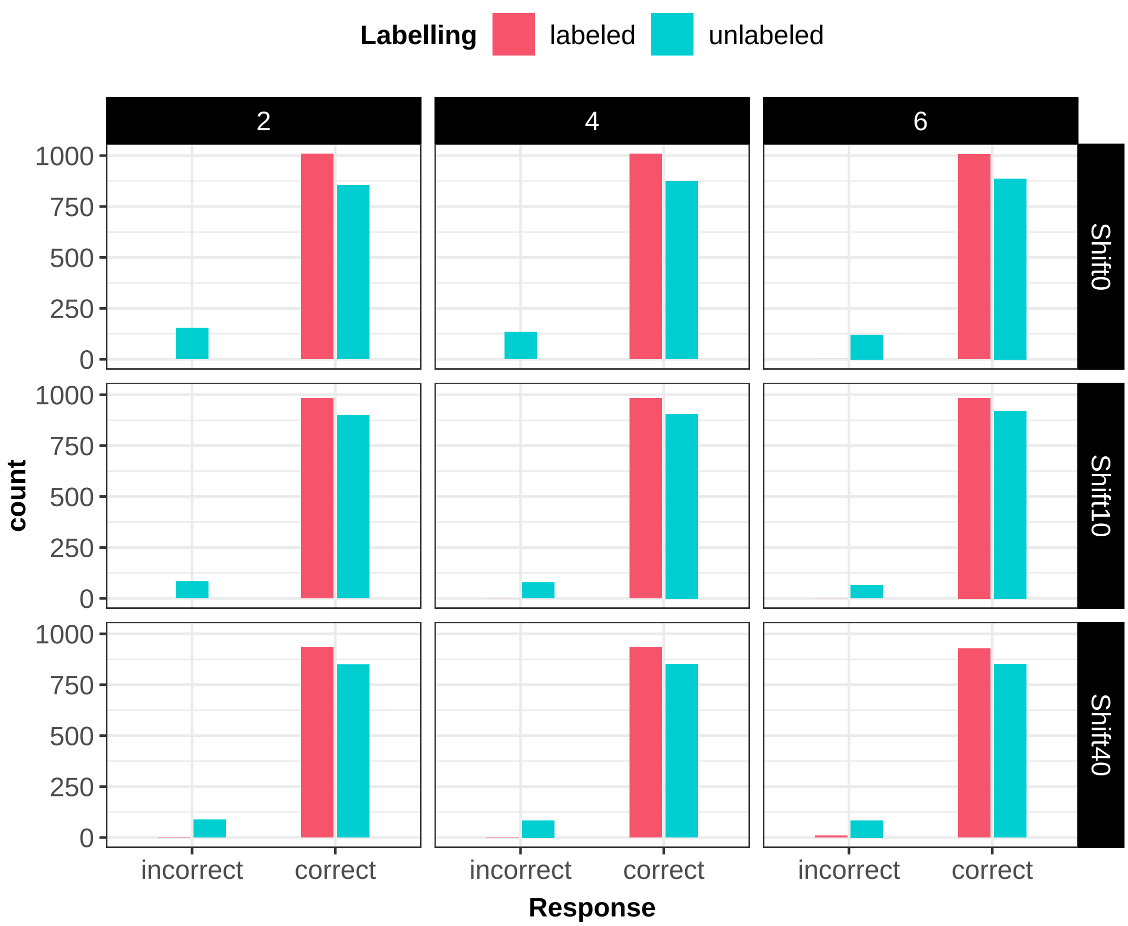

Figure 15

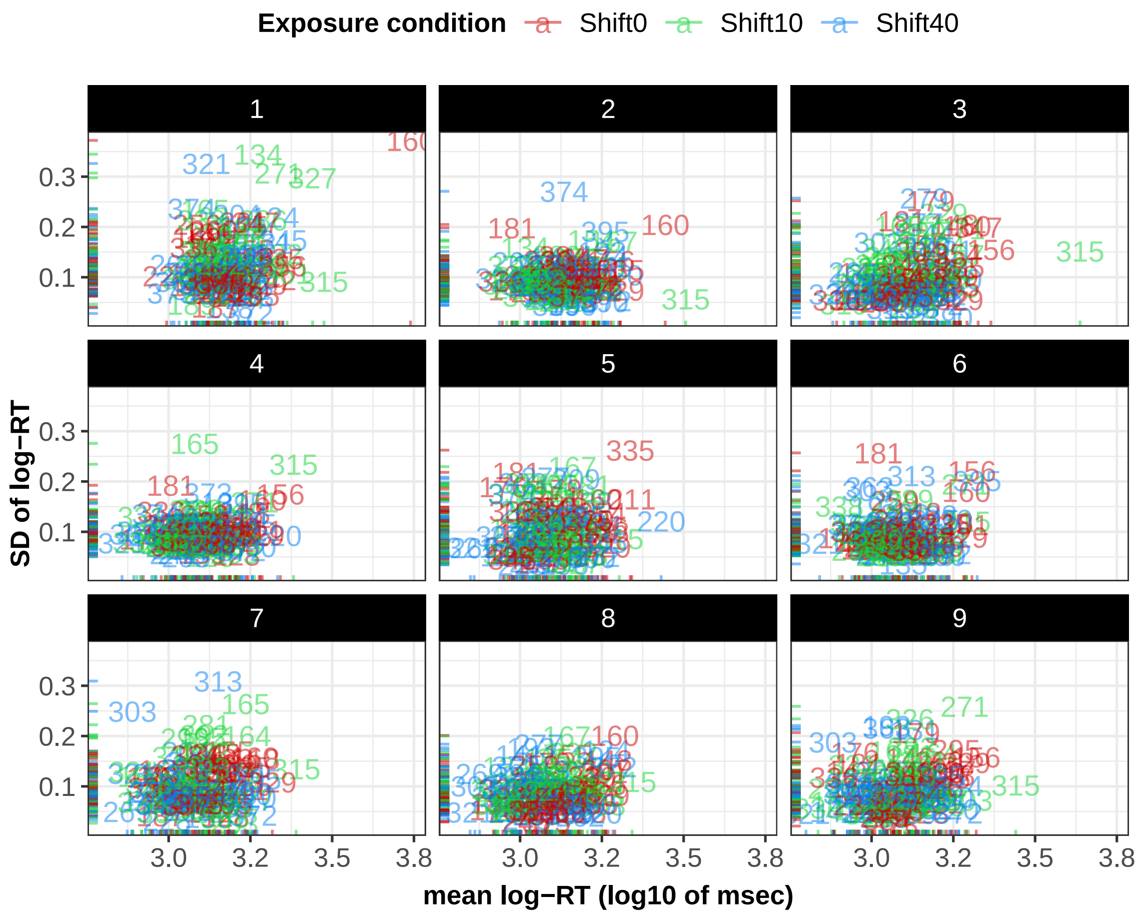

Figure 16

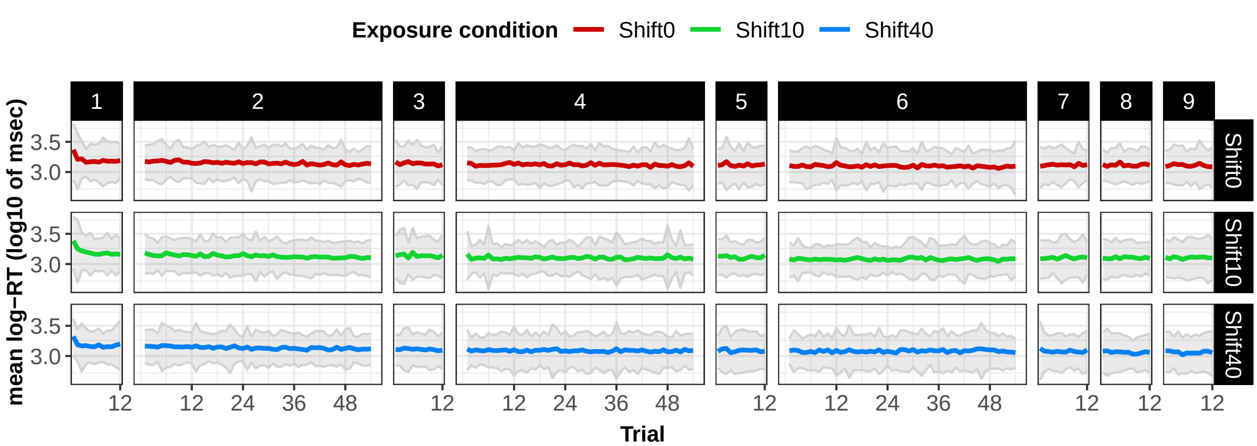

Figure 17

the product of the likelihood of the cue under the hypothesis that the talker produced “t”, and the prior probability of that cue. By using IOs trained solely on production data to predict categorization behaviour we avoid additional computational degrees of freedom and limit the risk of overfitting the model to the data thus reducing bias.

We filtered the database to /d/s and /t/s which gave 92 talkers (4x male and 4x female), each with a minimum of 25 tokens. We then fit ideal observers to each talker under different hypotheses of distributional learning [and evaluated their respective goodness-of-fit to the human data]. In total we fit x IOs to represent the different hypotheses about listeners’ implicit knowledge – models grouped by sex, grouped by sex and Predictions of the IO were obtained using talker-normalized category statistics for /d/ and /t/ from (X. Xie et al., 2022) based on data from (**chodroff2017?**), perceptual noise estimates for VOT from (Kronrod et al., 2016), and a lapse rate identical to the psychometric model estimate.

§5 Session Info

```
## - Session info ----------------------------------------------------------------------------
##  setting  value
##  version  R version 4.1.3 (2022-03-10)
##  os       macOS Big Sur/Monterey 10.16
##  system   x86_64, darwin17.0
##  ui       X11
##  language (EN)
##  collate  en_US.UTF-8
##  ctype    en_US.UTF-8
##  tz       Europe/Stockholm
##  date     2023-04-10
##  pandoc   2.18 @ /Applications/RStudio.app/Contents/MacOS/quarto/bin/tools/ (via rmarkdown)
##
## - Packages --------------------------------------------------------------------------------
##  package        * version    date (UTC) lib source
```

```
## abind              1.4-5      2016-07-21 [1] CRAN (R 4.1.0)
## arrayhelpers       1.1-0      2020-02-04 [1] CRAN (R 4.1.0)
## assertthat       * 0.2.1      2019-03-21 [1] CRAN (R 4.1.0)
## av                 0.8.3      2023-02-05 [1] CRAN (R 4.1.2)
## backports          1.4.1      2021-12-13 [1] CRAN (R 4.1.0)
## base64enc          0.1-3      2015-07-28 [1] CRAN (R 4.1.0)
## bayesplot          1.10.0     2022-11-16 [1] CRAN (R 4.1.2)
## bayestestR         0.13.0     2022-09-18 [1] CRAN (R 4.1.2)
## bit                4.0.5      2022-11-15 [1] CRAN (R 4.1.2)
## bit64              4.0.5      2020-08-30 [1] CRAN (R 4.1.0)
## bookdown           0.33       2023-03-06 [1] CRAN (R 4.1.2)
## boot               1.3-28.1   2022-11-22 [1] CRAN (R 4.1.2)
## bridgesampling     1.1-2      2021-04-16 [1] CRAN (R 4.1.0)
## brms             * 2.18.0     2022-09-19 [1] CRAN (R 4.1.2)
## Brobdingnag        1.2-9      2022-10-19 [1] CRAN (R 4.1.2)
## broom              1.0.4      2023-03-11 [1] CRAN (R 4.1.2)
## broom.mixed      * 0.2.9.4    2022-04-17 [1] CRAN (R 4.1.2)
## cachem             1.0.7      2023-02-24 [1] CRAN (R 4.1.3)
## callr              3.7.3      2022-11-02 [1] CRAN (R 4.1.2)
## car                3.1-1      2022-10-19 [1] CRAN (R 4.1.2)
## carData            3.0-5      2022-01-06 [1] CRAN (R 4.1.2)
## cellranger         1.1.0      2016-07-27 [1] CRAN (R 4.1.0)
## checkmate          2.1.0      2022-04-21 [1] CRAN (R 4.1.2)
## class              7.3-20     2022-01-16 [1] CRAN (R 4.1.3)
## classInt           0.4-8      2022-09-29 [1] CRAN (R 4.1.2)
## cli                3.6.0      2023-01-09 [1] CRAN (R 4.1.2)
## cluster            2.1.4      2022-08-22 [1] CRAN (R 4.1.2)
## coda               0.19-4     2020-09-30 [1] CRAN (R 4.1.0)
## codetools          0.2-18     2020-11-04 [1] CRAN (R 4.1.3)
## colorspace         2.1-0      2023-01-23 [1] CRAN (R 4.1.2)
```

```
## colourpicker      1.2.0    2022-10-28 [1] CRAN (R 4.1.2)
## cowplot         * 1.1.1    2020-12-30 [1] CRAN (R 4.1.0)
## crayon            1.5.2    2022-09-29 [1] CRAN (R 4.1.2)
## crosstalk         1.2.0    2021-11-04 [1] CRAN (R 4.1.0)
## curl            * 4.3.3    2022-10-06 [1] CRAN (R 4.1.2)
## data.table        1.14.8   2023-02-17 [1] CRAN (R 4.1.2)
## datawizard        0.6.4    2022-11-19 [1] CRAN (R 4.1.2)
## DBI               1.1.3    2022-06-18 [1] CRAN (R 4.1.2)
## dbplyr            2.2.1    2022-06-27 [1] CRAN (R 4.1.2)
## deldir            1.0-6    2021-10-23 [1] CRAN (R 4.1.0)
## devtools          2.4.5    2022-10-11 [1] CRAN (R 4.1.2)
## digest            0.6.31   2022-12-11 [1] CRAN (R 4.1.2)
## diptest         * 0.76-0   2021-05-04 [1] CRAN (R 4.1.0)
## distributional    0.3.1    2022-09-02 [1] CRAN (R 4.1.2)
## dplyr           * 1.1.0    2023-01-29 [1] CRAN (R 4.1.2)
## DT                0.26     2022-10-19 [1] CRAN (R 4.1.2)
## dygraphs          1.1.1.6  2018-07-11 [1] CRAN (R 4.1.0)
## e1071             1.7-13   2023-02-01 [1] CRAN (R 4.1.2)
## effectsize        0.8.2    2022-10-31 [1] CRAN (R 4.1.2)
## ellipse           0.4.3    2022-05-31 [1] CRAN (R 4.1.2)
## ellipsis          0.3.2    2021-04-29 [1] CRAN (R 4.1.0)
## emmeans           1.8.2    2022-10-27 [1] CRAN (R 4.1.2)
## estimability      1.4.1    2022-08-05 [1] CRAN (R 4.1.2)
## evaluate          0.20     2023-01-17 [1] CRAN (R 4.1.2)
## extraDistr        1.9.1    2020-09-07 [1] CRAN (R 4.1.0)
## fansi             1.0.4    2023-01-22 [1] CRAN (R 4.1.2)
## farver            2.1.1    2022-07-06 [1] CRAN (R 4.1.2)
## fastmap           1.1.1    2023-02-24 [1] CRAN (R 4.1.3)
## forcats         * 1.0.0    2023-01-29 [1] CRAN (R 4.1.2)
## foreach           1.5.2    2022-02-02 [1] CRAN (R 4.1.2)
```

```
## foreign          0.8-83    2022-09-28 [1] CRAN (R 4.1.2)
## Formula          1.2-5     2023-02-24 [1] CRAN (R 4.1.3)
## fs               1.6.1     2023-02-06 [1] CRAN (R 4.1.3)
## furrr            0.3.1     2022-08-15 [1] CRAN (R 4.1.2)
## future           1.29.0    2022-11-06 [1] CRAN (R 4.1.2)
## gargle           1.2.1     2022-09-08 [1] CRAN (R 4.1.2)
## generics         0.1.3     2022-07-05 [1] CRAN (R 4.1.2)
## gganimate        1.0.8     2022-09-08 [1] CRAN (R 4.1.2)
## ggdist           3.2.1     2023-01-18 [1] CRAN (R 4.1.2)
## ggforce          0.4.1     2022-10-04 [1] CRAN (R 4.1.2)
## ggnewscale       0.4.8     2022-10-06 [1] CRAN (R 4.1.2)
## ggplot2        * 3.4.1     2023-02-10 [1] CRAN (R 4.1.3)
## ggpubr           0.5.0     2022-11-16 [1] CRAN (R 4.1.2)
## ggrepel          0.9.2     2022-11-06 [1] CRAN (R 4.1.2)
## ggridges         0.5.4     2022-09-26 [1] CRAN (R 4.1.2)
## ggsignif         0.6.4     2022-10-13 [1] CRAN (R 4.1.2)
## ggstance       * 0.3.6     2022-11-16 [1] CRAN (R 4.1.2)
## gifski           1.6.6-1   2022-04-05 [1] CRAN (R 4.1.2)
## globals          0.16.2    2022-11-21 [1] CRAN (R 4.1.2)
## glue             1.6.2     2022-02-24 [1] CRAN (R 4.1.2)
## googledrive      2.0.0     2021-07-08 [1] CRAN (R 4.1.0)
## googlesheets4    1.0.1     2022-08-13 [1] CRAN (R 4.1.2)
## gridExtra        2.3       2017-09-09 [1] CRAN (R 4.1.0)
## gtable           0.3.1     2022-09-01 [1] CRAN (R 4.1.2)
## gtools           3.9.4     2022-11-27 [1] CRAN (R 4.1.2)
## haven            2.5.1     2022-08-22 [1] CRAN (R 4.1.2)
## HDInterval       0.2.4     2022-11-17 [1] CRAN (R 4.1.2)
## Hmisc            4.8-0     2023-02-09 [1] CRAN (R 4.1.2)
## hms              1.1.2     2022-08-19 [1] CRAN (R 4.1.2)
## htmlTable        2.4.1     2022-07-07 [1] CRAN (R 4.1.2)
```

```
## htmltools         0.5.4    2022-12-07 [1] CRAN (R 4.1.2)
## htmlwidgets       1.6.1    2023-01-07 [1] CRAN (R 4.1.2)
## httpuv            1.6.6    2022-09-08 [1] CRAN (R 4.1.2)
## httr              1.4.4    2022-08-17 [1] CRAN (R 4.1.2)
## igraph            1.3.5    2022-09-22 [1] CRAN (R 4.1.2)
## inline            0.3.19   2021-05-31 [1] CRAN (R 4.1.2)
## insight           0.18.8   2022-11-24 [1] CRAN (R 4.1.2)
## interp            1.1-3    2022-07-13 [1] CRAN (R 4.1.2)
## iterators         1.0.14   2022-02-05 [1] CRAN (R 4.1.2)
## jpeg              0.1-10   2022-11-29 [1] CRAN (R 4.1.2)
## jsonlite          1.8.4    2022-12-06 [1] CRAN (R 4.1.2)
## kableExtra      * 1.3.4    2021-02-20 [1] CRAN (R 4.1.2)
## KernSmooth        2.23-20  2021-05-03 [1] CRAN (R 4.1.3)
## knitr             1.42     2023-01-25 [1] CRAN (R 4.1.2)
## labeling          0.4.2    2020-10-20 [1] CRAN (R 4.1.0)
## LaplacesDemon     16.1.6   2021-07-09 [1] CRAN (R 4.1.0)
## later             1.3.0    2021-08-18 [1] CRAN (R 4.1.0)
## latexdiffr      * 0.1.0    2021-05-03 [1] CRAN (R 4.1.0)
## lattice           0.20-45  2021-09-22 [1] CRAN (R 4.1.3)
## latticeExtra      0.6-30   2022-07-04 [1] CRAN (R 4.1.2)
## lazyeval          0.2.2    2019-03-15 [1] CRAN (R 4.1.0)
## lifecycle         1.0.3    2022-10-07 [1] CRAN (R 4.1.2)
## linguisticsdown * 1.2.0    2019-03-01 [1] CRAN (R 4.1.0)
## listenv           0.8.0    2019-12-05 [1] CRAN (R 4.1.0)
## lme4            * 1.1-31   2022-11-01 [1] CRAN (R 4.1.2)
## lmerTest          3.1-3    2020-10-23 [1] CRAN (R 4.1.0)
## loo               2.5.1    2022-03-24 [1] CRAN (R 4.1.2)
## lpSolve           5.6.18   2023-02-01 [1] CRAN (R 4.1.2)
## lubridate         1.9.0    2022-11-06 [1] CRAN (R 4.1.2)
## magick          * 2.7.3    2021-08-18 [1] CRAN (R 4.1.0)
```

```
## magrittr        * 2.0.3      2022-03-30 [1] CRAN (R 4.1.2)
## markdown          1.4        2022-11-16 [1] CRAN (R 4.1.2)
## MASS            * 7.3-58.2   2023-01-23 [1] CRAN (R 4.1.2)
## Matrix          * 1.5-1      2022-09-13 [1] CRAN (R 4.1.2)
## matrixStats       0.63.0     2022-11-18 [1] CRAN (R 4.1.2)
## memoise           2.0.1      2021-11-26 [1] CRAN (R 4.1.0)
## mime              0.12       2021-09-28 [1] CRAN (R 4.1.0)
## miniUI            0.1.1.1    2018-05-18 [1] CRAN (R 4.1.0)
## minqa             1.2.5      2022-10-19 [1] CRAN (R 4.1.2)
## modelr            0.1.10     2022-11-11 [1] CRAN (R 4.1.2)
## multcomp          1.4-20     2022-08-07 [1] CRAN (R 4.1.2)
## munsell           0.5.0      2018-06-12 [1] CRAN (R 4.1.0)
## MVBeliefUpdatr  * 0.0.1.0002 2023-02-25 [1] Github (hlplab/MVBeliefUpdatr@2f7690c)
## mvtnorm           1.1-3      2021-10-08 [1] CRAN (R 4.1.0)
## nlme              3.1-160    2022-10-10 [1] CRAN (R 4.1.2)
## nloptr            2.0.3      2022-05-26 [1] CRAN (R 4.1.2)
## nnet              7.3-18     2022-09-28 [1] CRAN (R 4.1.2)
## numDeriv          2016.8-1.1 2019-06-06 [1] CRAN (R 4.1.0)
## pander            0.6.5      2022-03-18 [1] CRAN (R 4.1.2)
## papaja          * 0.1.1.9001 2023-03-21 [1] Github (crsh/papaja@c39033a)
## parallelly        1.32.1     2022-07-21 [1] CRAN (R 4.1.2)
## parameters        0.20.0     2022-11-21 [1] CRAN (R 4.1.2)
## patchwork       * 1.1.2      2022-08-19 [1] CRAN (R 4.1.2)
## phonR           * 1.0-7      2016-08-25 [1] CRAN (R 4.1.0)
## pillar            1.8.1      2022-08-19 [1] CRAN (R 4.1.2)
## pkgbuild          1.4.0      2022-11-27 [1] CRAN (R 4.1.2)
## pkgconfig         2.0.3      2019-09-22 [1] CRAN (R 4.1.0)
## pkgload           1.3.2      2022-11-16 [1] CRAN (R 4.1.2)
## plotly            4.10.1     2022-11-07 [1] CRAN (R 4.1.2)
## plyr              1.8.8      2022-11-11 [1] CRAN (R 4.1.2)
```

```
## png              0.1-8    2022-11-29 [1] CRAN (R 4.1.3)
## polyclip         1.10-4   2022-10-20 [1] CRAN (R 4.1.2)
## posterior      * 1.4.0    2023-02-22 [1] CRAN (R 4.1.2)
## prettyunits      1.1.1    2020-01-24 [1] CRAN (R 4.1.0)
## processx         3.8.0    2022-10-26 [1] CRAN (R 4.1.2)
## profvis          0.3.7    2020-11-02 [1] CRAN (R 4.1.0)
## progress         1.2.2    2019-05-16 [1] CRAN (R 4.1.0)
## promises         1.2.0.1  2021-02-11 [1] CRAN (R 4.1.0)
## proxy            0.4-27   2022-06-09 [1] CRAN (R 4.1.2)
## ps               1.7.2    2022-10-26 [1] CRAN (R 4.1.2)
## purrr          * 1.0.1    2023-01-10 [1] CRAN (R 4.1.2)
## R6               2.5.1    2021-08-19 [1] CRAN (R 4.1.0)
## rbibutils        2.2.13   2023-01-13 [1] CRAN (R 4.1.2)
## RColorBrewer     1.1-3    2022-04-03 [1] CRAN (R 4.1.2)
## Rcpp           * 1.0.10   2023-01-22 [1] CRAN (R 4.1.2)
## RcppParallel     5.1.6    2023-01-09 [1] CRAN (R 4.1.2)
## Rdpack           2.4      2022-07-20 [1] CRAN (R 4.1.2)
## readr          * 2.1.3    2022-10-01 [1] CRAN (R 4.1.2)
## readxl           1.4.1    2022-08-17 [1] CRAN (R 4.1.2)
## remotes          2.4.2    2021-11-30 [1] CRAN (R 4.1.0)
## reprex           2.0.2    2022-08-17 [1] CRAN (R 4.1.2)
## reshape2         1.4.4    2020-04-09 [1] CRAN (R 4.1.0)
## rlang          * 1.1.0    2023-03-14 [1] CRAN (R 4.1.2)
## rmarkdown        2.20     2023-01-19 [1] CRAN (R 4.1.2)
## rpart            4.1.19   2022-10-21 [1] CRAN (R 4.1.2)
## rsample        * 1.1.1    2022-12-07 [1] CRAN (R 4.1.2)
## rstan            2.21.8   2023-01-17 [1] CRAN (R 4.1.2)
## rstantools       2.2.0    2022-04-08 [1] CRAN (R 4.1.2)
## rstatix          0.7.1    2022-11-09 [1] CRAN (R 4.1.2)
## rstudioapi       0.14     2022-08-22 [1] CRAN (R 4.1.2)
```

```
## rvest           1.0.3    2022-08-19 [1] CRAN (R 4.1.2)
## sandwich        3.0-2    2022-06-15 [1] CRAN (R 4.1.2)
## scales          1.2.1    2022-08-20 [1] CRAN (R 4.1.2)
## sessioninfo     1.2.2    2021-12-06 [1] CRAN (R 4.1.0)
## sf              1.0-9    2022-11-08 [1] CRAN (R 4.1.2)
## shiny           1.7.3    2022-10-25 [1] CRAN (R 4.1.2)
## shinyjs         2.1.0    2021-12-23 [1] CRAN (R 4.1.0)
## shinystan       2.6.0    2022-03-03 [1] CRAN (R 4.1.2)
## shinythemes     1.2.0    2021-01-25 [1] CRAN (R 4.1.0)
## StanHeaders     2.21.0-7 2020-12-17 [1] CRAN (R 4.1.0)
## stringi         1.7.12   2023-01-11 [1] CRAN (R 4.1.2)
## stringr       * 1.5.0    2022-12-02 [1] CRAN (R 4.1.2)
## survival        3.4-0    2022-08-09 [1] CRAN (R 4.1.2)
## svglite         2.1.0    2022-02-03 [1] CRAN (R 4.1.2)
## svUnit          1.0.6    2021-04-19 [1] CRAN (R 4.1.0)
## systemfonts     1.0.4    2022-02-11 [1] CRAN (R 4.1.2)
## tensorA         0.36.2   2020-11-19 [1] CRAN (R 4.1.0)
## TH.data         1.1-1    2022-04-26 [1] CRAN (R 4.1.2)
## threejs         0.3.3    2020-01-21 [1] CRAN (R 4.1.0)
## tibble        * 3.2.1    2023-03-20 [1] CRAN (R 4.1.3)
## tidybayes     * 3.0.3    2023-02-04 [1] CRAN (R 4.1.2)
## tidyr         * 1.3.0    2023-01-24 [1] CRAN (R 4.1.2)
## tidyselect      1.2.0    2022-10-10 [1] CRAN (R 4.1.2)
## tidyverse     * 1.3.2    2022-07-18 [1] CRAN (R 4.1.2)
## timechange      0.1.1    2022-11-04 [1] CRAN (R 4.1.2)
## tinylabels    * 0.2.3    2022-02-06 [1] CRAN (R 4.1.2)
## transformr      0.1.4    2022-08-18 [1] CRAN (R 4.1.2)
## tufte           0.12     2022-01-27 [1] CRAN (R 4.1.2)
## tweenr          2.0.2    2022-09-06 [1] CRAN (R 4.1.2)
## tzdb            0.3.0    2022-03-28 [1] CRAN (R 4.1.2)
```

```
## units              0.8-1    2022-12-10 [1] CRAN (R 4.1.2)
## urlchecker         1.0.1    2021-11-30 [1] CRAN (R 4.1.0)
## usethis            2.1.6    2022-05-25 [1] CRAN (R 4.1.2)
## utf8               1.2.3    2023-01-31 [1] CRAN (R 4.1.2)
## vctrs              0.6.0    2023-03-16 [1] CRAN (R 4.1.3)
## viridis            0.6.2    2021-10-13 [1] CRAN (R 4.1.0)
## viridisLite        0.4.1    2022-08-22 [1] CRAN (R 4.1.2)
## vroom              1.6.0    2022-09-30 [1] CRAN (R 4.1.2)
## webshot          * 0.5.4    2022-09-26 [1] CRAN (R 4.1.2)
## withr              2.5.0    2022-03-03 [1] CRAN (R 4.1.2)
## xfun               0.37     2023-01-31 [1] CRAN (R 4.1.2)
## xml2               1.3.3    2021-11-30 [1] CRAN (R 4.1.0)
## xtable             1.8-4    2019-04-21 [1] CRAN (R 4.1.0)
## xts                0.12.2   2022-10-16 [1] CRAN (R 4.1.2)
## yaml               2.3.7    2023-01-23 [1] CRAN (R 4.1.2)
## zoo                1.8-11   2022-09-17 [1] CRAN (R 4.1.2)
##
## [1] /Library/Frameworks/R.framework/Versions/4.1/Resources/library
##
## ------------------------------------------------------------------------------------------
```

Table 1
Population estimates

term	estimate	std.error	conf.low
mu2__(Intercept)	-3.10	1.12	-5.02
theta1__(Intercept)	-4.78	0.36	-5.60
mu2__VOT__gs	16.93	1.78	13.57
mu2__Condition.Exposure__Shift0vs.Shift10	-1.05	0.79	-2.52
mu2__Condition.Exposure__Shift10vs.Shift40	-2.40	0.85	-4.04
mu2__Block__Block1vs.Block3	0.03	1.06	-1.92
mu2__Block__Block3vs.Block5	0.09	0.86	-1.60
mu2__Block__Block5vs.Block7	-0.05	0.88	-1.68
mu2__Block__Block7vs.Block8	0.05	0.62	-1.11
mu2__Block__Block8vs.Block9	0.10	0.80	-1.58
mu2__VOT__gs:Condition.Exposure__Shift0vs.Shift10	1.17	1.87	-2.51
mu2__VOT__gs:Condition.Exposure__Shift10vs.Shift40	-0.81	1.81	-4.31
mu2__VOT__gs:Block__Block1vs.Block3	0.29	2.26	-4.29
mu2__VOT__gs:Block__Block3vs.Block5	-1.69	1.97	-5.80
mu2__VOT__gs:Block__Block5vs.Block7	0.94	1.68	-2.31
mu2__VOT__gs:Block__Block7vs.Block8	-0.22	1.77	-3.73
mu2__VOT__gs:Block__Block8vs.Block9	1.69	1.91	-2.07
mu2__Condition.Exposure__Shift0vs.Shift10:Block__Block1vs.Block3	-1.41	1.23	-3.77
mu2__Condition.Exposure__Shift10vs.Shift40:Block__Block1vs.Block3	-2.21	1.32	-4.89
mu2__Condition.Exposure__Shift0vs.Shift10:Block__Block3vs.Block5	1.01	1.35	-1.65
mu2__Condition.Exposure__Shift10vs.Shift40:Block__Block3vs.Block5	-1.68	1.32	-4.27
mu2__Condition.Exposure__Shift0vs.Shift10:Block__Block5vs.Block7	-0.12	1.24	-2.55
mu2__Condition.Exposure__Shift10vs.Shift40:Block__Block5vs.Block7	-0.30	1.51	-3.17
mu2__Condition.Exposure__Shift0vs.Shift10:Block__Block7vs.Block8	-0.34	0.93	-2.21
mu2__Condition.Exposure__Shift10vs.Shift40:Block__Block7vs.Block8	1.15	1.24	-1.30
mu2__Condition.Exposure__Shift0vs.Shift10:Block__Block8vs.Block9	1.28	1.02	-0.61
mu2__Condition.Exposure__Shift10vs.Shift40:Block__Block8vs.Block9	0.96	1.23	-1.49
mu2__VOT__gs:Condition.Exposure__Shift0vs.Shift10:Block__Block1vs.Block3	4.81	3.97	-2.88
mu2__VOT__gs:Condition.Exposure__Shift10vs.Shift40:Block__Block1vs.Block3	-5.83	3.59	-12.74
mu2__VOT__gs:Condition.Exposure__Shift0vs.Shift10:Block__Block3vs.Block5	-3.57	3.28	-10.15
mu2__VOT__gs:Condition.Exposure__Shift10vs.Shift40:Block__Block3vs.Block5	4.62	3.31	-1.72
mu2__VOT__gs:Condition.Exposure__Shift0vs.Shift10:Block__Block5vs.Block7	1.18	3.03	-4.65
mu2__VOT__gs:Condition.Exposure__Shift10vs.Shift40:Block__Block5vs.Block7	0.40	3.35	-6.08
mu2__VOT__gs:Condition.Exposure__Shift0vs.Shift10:Block__Block7vs.Block8	-0.98	3.13	-7.49
mu2__VOT__gs:Condition.Exposure__Shift10vs.Shift40:Block__Block7vs.Block8	0.09	3.25	-6.24
mu2__VOT__gs:Condition.Exposure__Shift0vs.Shift10:Block__Block8vs.Block9	-3.85	3.28	-10.42
mu2__VOT__gs:Condition.Exposure__Shift10vs.Shift40:Block__Block8vs.Block9	-3.96	3.11	-10.23