Unravelling the time-course of listener adaptation to 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

Unravelling the time-course of listener adaptation to 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 vary in the way they realise linguistic categories. Yet, listeners who share a common language background typically cope with talker variability without difficulty. In scenarios where a talker produces those categories in an unexpected and unfamiliar way comprehension may become a real challenge. It has been shown, however that brief exposure to unfamiliar accents 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 robust comprehension of speech in spite of this variability (the classic "lack of invariance" problem) remains the a longstanding question in speech perception research.

In the past two decades the hypothesis that listeners overcome the lack of invariance by learning the distributions of acoustic cue-to-linguistic category mappings has gained considerable influence in contemporary approaches to studying this problem. A growing number of studies have demonstrated that changes in listener behaviour through the course of a short experiment align qualitatively with the statistics of exposure stimuli Theodore & Monto (2019).

For example when listeners are tasked with identifying word pairs contrasted by the voice-onset-time (VOT) cue (e.g. "beach-peach") they would respond with greater uncertainty (less steep categorisation functions) if tokens of these words had VOTs are sampled from a bi-modal distribution with large variances Theodore & Monto (2019). In contrast, listeners who heard the same minimal-pair words with VOT values sampled from distributions with narrow variances gave more categorical responses, producing steeper categorisation.

Impetus for such investigations follows from the hypothesis that listeners hold implicit knowledge of how talkers should sound in a given linguistic encounter. Drawing from this implicit ### ### In a follow-up 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 by varying degrees, as well as leftwards, from the expected mean values of a typical L1 US English. With this manipulation of means they investigated 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.

–WHAT'S THE MOTIVATION FOR THIS WORK– 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. In the computation of this process, each additional trial of exposure should cause listeners to move progressively closer to the target classification boundary implied by the exposure distribution. The overall design of the study was 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.

POINTS-TO-MAKE ___Most of the work has focused on the outcome of exposure. - Qualitatively, we know that exposing listeners to different distributions produces changes in categorisation behaviour towards the direction of the shifts. - A stronger test for the computational framework is needed. - The ideal adapter framework makes specific predictions about rational speech perception. For example, listeners' integrate the exposure with their prior knowledge and infer the cue-category distributions of a talker. Listeners hold implicit beliefs or expectations about the distributions of cues which they bring to an encounter. - The strength of these beliefs has bearing on listener propensity to adapt to a new talker – the stronger the prior beliefs the longer it takes to adapt. Listeners' strengths in prior beliefs about the means and variances are represented by parameters in the computational model. Listener behaviour observed collectively, thus far which speaks to this framework of thinking should by now be able to indicate roughly what those parameter values are. But it looks like those parameters are biased by the length of exposure and the outcome during experiments. No one has been able to confront this issue of very quick but limited adaptation which can't be solved by giving more exposure trials. - Do the data indicate that distributional learning also includes Bayesian belief-updating? - How do we distinguish the results from normalization accounts which can also explain adaptation?

–ECOLOGICAL VALIDITY OF PARADIGM– 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).

1.1 Methods

1.1.1 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.

1.1.2 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 from New Hampshire, judged to have a "general American" 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). 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. The full procedure is described in the supplementary information (SI, **??**).

We also 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 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.

Prior to creating the three exposure conditions of the experiment, we ran a norming experiment to test US-L1 listeners' perception of our stimuli and to determine a baseline categorisation boundary for this talker. While it is normal and acceptable practice to set the baseline by taking population estimates of mean values from past studies on stops, we reasoned that such estimates were highly variable and therefore aimed to obtained a more accurate estimation of how L1-US English listeners perceived the speech of our talker. To anticipate the outcome, we eventually discovered that the classification boundary from norming underestimated the boundary fitted to our participants' classification in the initial test block. This placed two of our exposure conditions slightly leftwards of participants' initial perceptual boundary. This outcome however does not impinge on the conclusions drawn from this study []

The other purpose of the norming experiment was to detect possible anomalous features present in our stimuli (for e.g. if it would elicit unusual categorisation behaviour or whether certain minimal-pairs had an exaggerated effect on categorisation). For the norming experiment the VOT continua employed 24 VOT steps ranging from -100ms VOT to +130ms (-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 found VOT to have the expected effect on the proportion of "t"-responses, i.e. higher VOTs elicited greater "t"-responses and that the word-pairs did not differ substantially from each other. The results and analysis of the norming experiment are reported in full in section **??**.

A subset of the materials were used to generate the three exposure conditions; 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 distinct as possible.

We employed a multi-block exposure-test design 1 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 minimise distortion of the intended distribution to be presented by the end of the exposure phase. After the final exposure block we tripled the number of test blocks to increase the statistical power to detect exposure induced behavioural changes.

The conditions were created by first generating a baseline distribution (+0ms shift) and then shifting that distribution by +10ms and by +40ms towards the right of the VOT continuum to create the remaining two conditions.

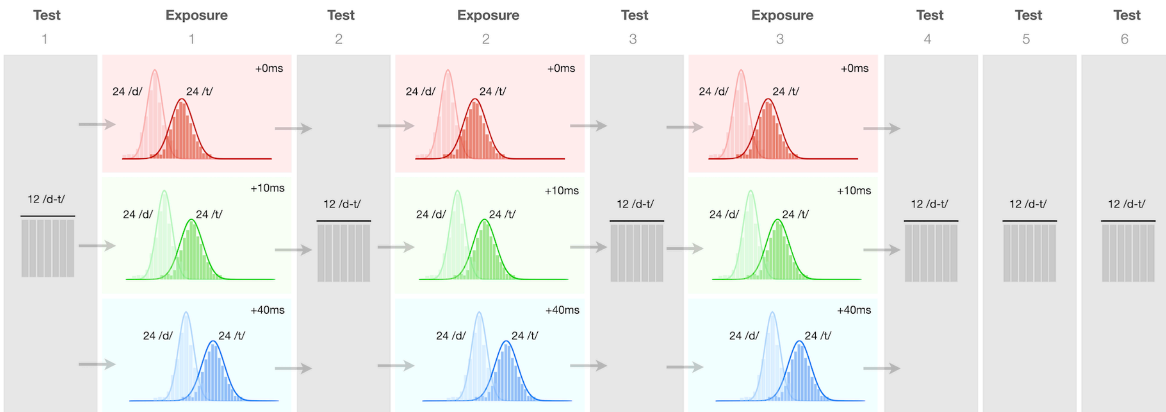

Figure 1. Experiment 2 multi-block design. Test blocks in grey comprised identical stimuli within and between conditions

To construct the +0ms shift exposure distribution we first computed the point of subjective equality (PSE) from the perceptual component of the fitted psychometric function of listener responses in the norming experiment. The PSE corresponds to the VOT duration that was perceived as most ambiguous across all participants during norming (i.e. the stimulus that on average, elicited equal chance of being categorised as /d/ or /t/) thus marking the categorical boundary. From a distributional perspective the PSE is where the likelihoods of both categories intersect and have equal probability 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 category likelihoods that could intersect at this value, we set the variances of the /d/ (80ms) and /t/ (270ms categories based on parameter estimates (X. Xie, Jaeger, and Kurumada (2022)) obtained from the production database of word-initial stops in Chodroff and Wilson (2017). To each variance value we added 80ms following (Kronrod, Coppess, and Feldman (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 mean for /d/ and /t/ estimated from 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 further detail on this procedure).

The distributional make up was determined through a process of sampling tokens from a discretised normal distribution with values rounded to the nearest multiple of 5 integer (available through the `extraDistr` package in R). For each exposure block 8 VOT tokens per minimal word pair were sampled from discrete normal distributions of each category of the +0ms condition, giving 24 /d/ and 24 /t/ items (48 critical trials) per block. The sampled distributions of VOT tokens were increased by a margin of +10ms and +40 ms to create the remaining two conditions 2. Additionally, each exposure block contained 2 instances of 3 catch items, giving 6 catch trials per block. These catch trials were recordings of the words, “flare”, “share”, or “rare”, presented in the same manner as critical trials but clearly distinguishable. They served as a check on participant attention during the experiment. 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*.

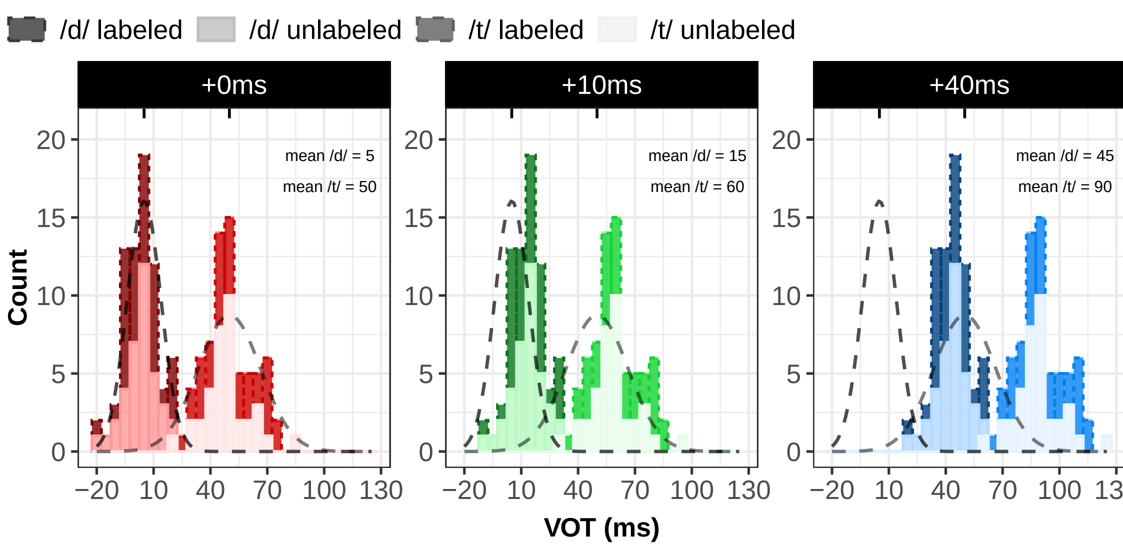

Figure 2

1.1.3 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.

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 **??**. At 1000ms from trial onset, the fixation dot would turn bright green and participants had to click on the dot to play the recording. Participants responded by clicking on the word they heard and the next trial would begin. The placement of the word presentations were counter-balanced across participants.

Participants underwent 234 trials which included 6 catch trials in each exposure block (18 in total). Since these recordings were easily distinguishable, they served as a check on participant attention throughout the experiment. 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 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.

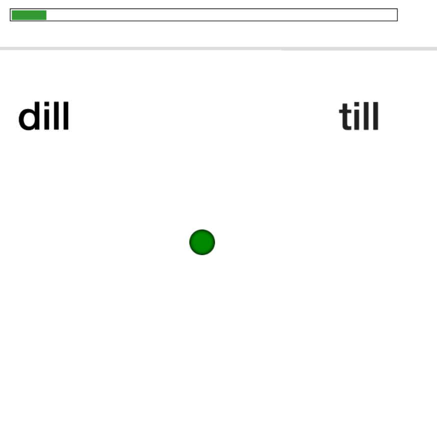

Figure 3. Example trial display. The words were displayed 500ms after trial onset. The green button would turn bright green signalling participants to click on the dot to play the recording.

1.1.4 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 labelled 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 below the empirical mean of the /d/ category and +20ms above the empirical mean of the /t/ category. These means were obtained from the production data estimates by X. Xie et al. (2022).

1.1.5 Analysis approach

1.2 Results

1.3 Regression analysis

The regression analysis addresses several questions: 1. Do listeners change their categorization behaviour in the direction predicted by their respective exposure distributions? 2. At what stage in the experiment did the behavioural change first emerge? 3. Are differences in categorisation behaviour in proportion to the difference between the exposure conditions? 4. Do the differences between exposure conditions diminish with repeated testing and without intermittent exposure?

We fit a Bayesian mixed-effects psychometric models to participants' categorization responses on critical test trials (e.g., **prins2011?**). We are primarily interested in the changes in categorization behaviour between test blocks which are presumed to be a consequence of the input from preceding exposure blocks however we fit a separate regression model for exposure in order to visualise participant behaviour during exposure.

The psychometric model is essentially an extension of mixed-effects logistic regression that also takes into account attentional lapses. Ignoring attentional lapses–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.

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.

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. We assumed a uniform bias in participant responses, that is, on trials where the participant was not paying attention he would respond “t” half the time and fitted a population-level intercept for the lapse rate. Random effects for the lapsing model and lapse rates were not fitted to reduce the number of parameters and to facilitate model convergence.

1.4 Behavioral results

```
## Hypothesis Tests for class b:
##          Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio Post.Prob Star
## 1 (mu2_VOT_gs) > 0       16       1.7       13       18       7999         1    *
## ---
## 'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.
## '*': For one-sided hypotheses, the posterior probability exceeds 95%;
## for two-sided hypotheses, the value tested against lies outside the 95%-CI.
## Posterior probabilities of point hypotheses assume equal prior probabilities.
```

Figure @**??**fig plot-fit-intercept-slope) summarizes participants’ fitted categorization functions across the different test blocks. A first point to note is the average categorization functions of the respective conditions before exposure to the talker. As depicted in the first panel, the average categorization functions converge on the same boundary or PSE (45ms, 95% QI = 36ms – 55ms) which suggests that the three exposure groups largely had similar expectations about the cue distribution corresponding to /d/ and /t/ for this type of talker.

There was a main effect of VOT $\hat{\beta} = 15.7$ 95%-CI: 12.5 to 19.2; Bayes factor: 7,999 90%-CI : 13.15 to 18.4; 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: -3.8 to -1.1; Bayes factor: 443.44 90%-CI : -3.54 to -1.36 reduction in log-odds) was *larger* than the difference between the +10 and +0 conditions ($\hat{\beta} =$ -1 95%-CI: -2.8 to 0.7; Bayes factor: 9.24 90%-CI : -2.24 to 0.3 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: -3.5 to 0.6; Bayes factor: 13.52 90%-CI : -3.06 to 0.2). 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 ($\hat{\beta} =$ -2.1 95%-CI: -4.4 to 0.2; Bayes factor: 27.78 90%-CI : -3.89 to -0.23).

In test block 3, the difference in average log-odds between conditions +0ms and +10ms, relative to test 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 ($\hat{\beta} =$ 0.8 95%-CI: -1.8 to 3.4; Bayes factor: 3.99 90%-CI : -1.11 to 2.78). 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 block-by-block differences between the +40ms and +10ms conditions, these 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 achieved their maximal shifts by block 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 in 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.

**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 predecessor, we also find that participants had a greater propensity to shift their categorisations rightwards towards higher VOT values rather than leftwards towards lower VOT values as the +40ms group showed the widest deviation from the baseline.

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. 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.

```
## Warning in tidy.brmsfit(fit_mix_exposure, effects = "fixed"): some parameter names contain u

## # A tibble: 19 x 5
##    term                                                                               estimate std
##    <chr>                                                                                 <dbl>
##  1 mu2_(Intercept)                                                                       -1.33
```

```
## 2 theta1_(Intercept)                                                          -6.86
## 3 mu2_VOT_gs                                                                  22.3
## 4 mu2_Condition.Exposure_Shift10vs.Shift0                                     -1.27
## 5 mu2_Condition.Exposure_Shift40vs.Shift10                                    -3.84
## 6 mu2_Block_Exposure2vs.Exposure1                                              0.0634
## 7 mu2_Block_Exposure3vs.Exposure2                                              0.0820
## 8 mu2_VOT_gs:Condition.Exposure_Shift10vs.Shift0                              -0.585
## 9 mu2_VOT_gs:Condition.Exposure_Shift40vs.Shift10                              2.48
## 10 mu2_VOT_gs:Block_Exposure2vs.Exposure1                                      0.912
## 11 mu2_VOT_gs:Block_Exposure3vs.Exposure2                                      0.810
## 12 mu2_Condition.Exposure_Shift10vs.Shift0:Block_Exposure2vs.Exposure1        -0.120
## 13 mu2_Condition.Exposure_Shift40vs.Shift10:Block_Exposure2vs.Exposure1       -0.613
## 14 mu2_Condition.Exposure_Shift10vs.Shift0:Block_Exposure3vs.Exposure2         0.214
## 15 mu2_Condition.Exposure_Shift40vs.Shift10:Block_Exposure3vs.Exposure2       -0.791
## 16 mu2_VOT_gs:Condition.Exposure_Shift10vs.Shift0:Block_Exposure2vs.Exposure1   0.963
## 17 mu2_VOT_gs:Condition.Exposure_Shift40vs.Shift10:Block_Exposure2vs.Exposure1 -6.31
## 18 mu2_VOT_gs:Condition.Exposure_Shift10vs.Shift0:Block_Exposure3vs.Exposure2  -4.00
## 19 mu2_VOT_gs:Condition.Exposure_Shift40vs.Shift10:Block_Exposure3vs.Exposure2  2.70

## [1] "VOT mean: 42.165"
## [1] "VOT sd: 30.3259"
## [1] "mean VOT is 42.1650326797386 and SD is 30.3259185098252"
##   _Exposure2 vs. Exposure1 _Exposure3 vs. Exposure2
## 2                    -0.67                    -0.33
## 4                     0.33                    -0.33
## 6                     0.33                     0.67
```

```
## Warning in tidy.brmsfit(fit_mix_test_nested, effects = "fixed"): some parameter names conta
```

```
## Warning in tidy.brmsfit(fit_mix_exposure_nested, effects = "fixed"): some parameter names co
```

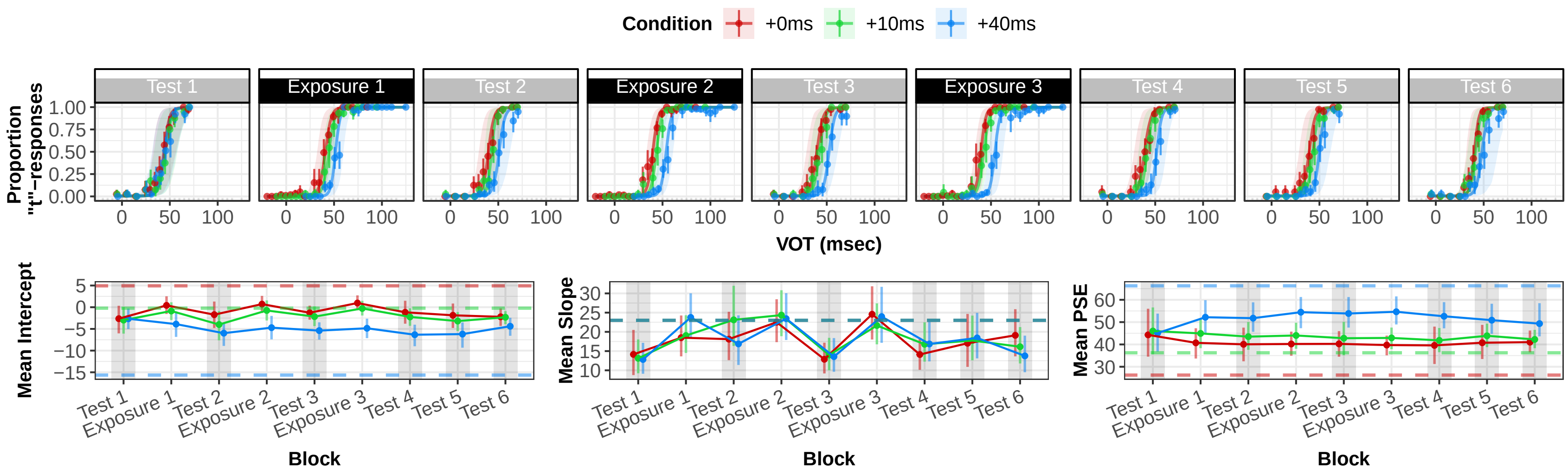

Figure 4. Fitted lapse-rate corrected psychometric plots by exposure condition; point ranges indicate the mean proportion of "t"-responses and their 95% bootstrapped CI (top row). Changes in intercepts, slopes and categorisation boundary (represented by the point-of-subjective-equality (PSE)) by block. Summary is of 8000 draws from the maximum *a posteriori* estimate. Points represent the mean of posterior draws and line ranges are the 95% quantile interval of all draws. Dashed lines show the predicted intercepts, slopes and PSEs by the ideal observers of the respective conditions that have perfectly learned the exposure distributions.

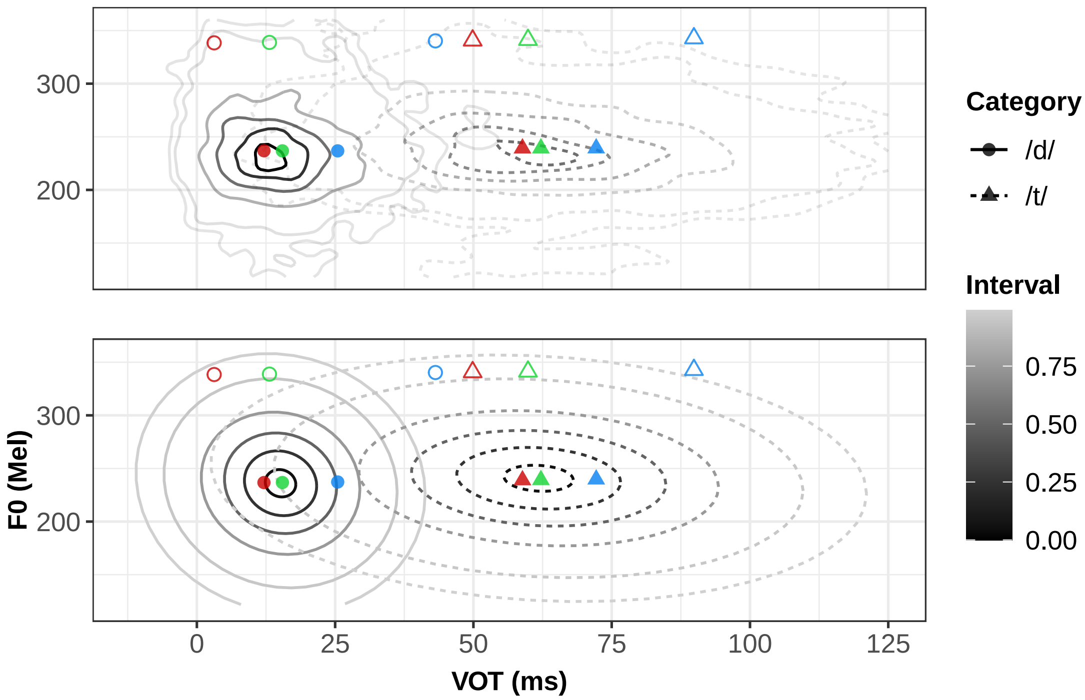
Category
/d/
/t/
Interval
0.75
0.50
0.25
0.00
300
200
F0 (Mel)
0
25
50
75
100
125
VOT (ms)

Figure 5

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).

2 General discussion

2.1 Methodological advances that can move the field forward

An example of a subsection.

3 References

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

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

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

Bradlow, A. R., & 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

Campitelli, E. (2022). *Ggnewscale: Multiple fill and colour scales in 'ggplot2'.* Retrieved from https://CRAN.R-project.org/package=ggnewscale

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.

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

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.

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/

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.

Hugh-Jones, D. (2021). *Latexdiffr: Diff 'rmarkdown' files using the 'latexdiff' utility*. Retrieved from https://CRAN.R-project.org/package=latexdiffr

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.

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. (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

Lüdecke, D. (2023). *sjPlot: Data visualization for statistics in social science.* Retrieved from https://CRAN.R-project.org/package=sjPlot

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

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

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

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

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

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

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., & 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

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/

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.

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 2.19.0; Bürkner, 2017, 2018, 2021), *broom.mixed* (Version 0.2.9.4; Bolker &

Robinson, 2022), *cowplot* (Version 1.1.1; Wilke, 2020), *curl* (Version 5.0.0; Ooms, 2022), *data.table* (Version 1.14.8; Dowle & Srinivasan, 2021), *diptest* (Version 0.76.0; Maechler, 2021), *dplyr* (Version 1.1.2; Wickham, François, Henry, & Müller, 2021), *forcats* (Version 1.0.0; Wickham, 2021a), *gganimate* (Version 1.0.8; Pedersen & Robinson, 2020), *ggdist* (Version 3.3.0; Kay, 2022a), *ggforce* (Version 0.4.1; Pedersen, 2022a), *ggnewscale* (Version 0.4.8; Campitelli, 2022), *ggplot2* (Version 3.4.2; Wickham, 2016), *ggpubr* (Version 0.6.0; Kassambara, 2020), *ggrepel* (Version 0.9.3; 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 1.1.33; Bates, Mächler, Bolker, & Walker, 2015), *lmerTest* (Version 3.1.3; Kuznetsova, Brockhoff, & Christensen, 2017), *lubridate* (Version 1.9.2; Grolemund & Wickham, 2011), *magick* (Version 2.7.4; Ooms, 2021), *magrittr* (Version 2.0.3; Bache & Wickham, 2020), *MASS* (Version 7.3.60; Venables & Ripley, 2002), *Matrix* (Version 1.5.1; Bates & Maechler, 2021), *modelr* (Version 0.1.11; 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 1.4.1; Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2021), *processx* (Version 3.8.1; 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 2.1.4; Wickham, Hester, & Bryan, 2021), *rlang* (Version 1.1.1; Henry & Wickham, 2021), *rsample* (Version 1.1.1; Frick et al., 2022), *scales* (Version 1.2.1; Wickham & Seidel, 2022), *sjPlot* (Version 2.8.14; Lüdecke, 2023), *stringr* (Version 1.5.0; Wickham, 2019b), *tibble* (Version 3.2.1; Müller & Wickham, 2021), *tidybayes* (Version 3.0.4; Kay, 2022b), *tidyr* (Version 1.3.0; Wickham, 2021b), *tidyverse* (Version 2.0.0; 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

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

Four /d-t/ minimal pairs (dill-till, din-tin, dim-tim, dip-tip) were recorded together with 20 filler words. These fillers were made up of 10 minimal or near minimal pairs with different sounds at onset. The word pairs were separated into two lists so that they would appear in separate blocks during recording. Each critical pair was repeated 8 times while the filler pairs were repeated 5 times. Word presentation was delivered with PsychoPy (**Peirce2019?**) and the presentation was controlled by the researcher from a computer located outside the recording room. The order of each block was randomised such that target words never appeared consecutively. The talker was instructed to speak clearly and confidently, and to maintain a consistent distance from the microphone.

§3.2 Annotation of audio stimuli

All critical pairs of the talker’s recordings were annotated. Durational, measurements of voicing lead, VOT, and vowel were taken in addition to the mean F0 of the first 25% of the vowel duration. Annotations were made with a combination of listening to the audio file and inspection of the waveform and spectrogram. The annotation boundaries were made according to the following principles:

- negative VOT (voicing during closure) – the start was marked as the first sign of periodicity in the waveform before closure release. The end was marked at the point of closure release

- VOT – start: the point of closure release. End: the beginning of clearly defined periodicity in the waveform and at the appearance of low frequency energy in the spectrogram.
- Vowel – start: the beginning of clearly defined periodicity in the waveform and at the appearance of low frequency energy in the spectrogram. End: if before a stop, when periodicity becomes irregular or at closure onset; if before a lateral, when formant transition approaches steady state; if before a nasal, when formants show a step-wise shift and when intensity shows a steep decline.
- F0 at vowel onset – the average pitch measurement estimated over the first 25% of the total vowel duration.

[INSERT EXAMPLE IMAGES]

§3.3 Synthesis of audio stimuli

The stimuli was created using the "progressive cutback and replacement method" by (Winn, 2020) implemented in Praat (Boersma & Weenink, 2022). This automates and greatly simplifies the process for generating highly natural sounding stimuli. Users of the script need only specify certain parameters to produce desired stimuli. Stimuli with pre-voicing were created separately from stimuli with positive VOT. This was because the script was not coded to automate the creation of tokens with pre-voicing that are natural sounding [1]. As such, the pre-voicing stimuli were created by prepending pre-voicing generated from naturally produced tokens (described below) that were edited with a separate process.

§3.4 Positive VOT tokens

For each minimal pair a continuum of 31 tokens was generated between 0ms and 150ms with a step-size of 5ms. A token of the voiced category from each pair was selected to be the base sound

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

file to make the continuum. All four minimal pair continua had an identical aspiration sound which was excised from one of the voiceless tokens produced by the talker.

While the main manipulation of the recordings was done on VOT we set the fundamental frequency (F0) to covary with VOT according to the natural correlation exhibited by our talker. The F0 values were predicted by regressing the talker's F0 measurements on VOT. Target F0 values for each token were then generated by setting the predicted F0 values of the end-point VOT tokens (0ms and 150ms) in the Praat script.

The vowel cut-back ratio was set at 0.33 which translates into a third of a ms vowel reduction for every 1ms of VOT. This ratio followed the estimated vowel duration-VOT trade-off for dip-tip minimal pair tokens reported in (**allenMiller?**). The maximum allowed vowel cut-back was 0.5ms to avoid the short vowel in **dip** becoming too short. Lastly, the rate of increase for aspiration intensity was kept at the default settings of the script.

§3.5 Pre-voicing tokens

Pre-voicing in 5ms increments were generated from a a clear pre-voicing waveform excised from a voiced token produced by the talker. To achieve a desired duration a duration factor is first computed and then converted with the "lengthen (overlap-add)" function in Praat. For example, if the desired amount of prevoicing was 50ms then the duration factor would be 50ms/length of the original pre-voicing sample. Each pre-voicing step is then prepended to a token with 0ms VOT. Each of these 0ms tokens was generated with Winn (2020) Praat script by manually entering the expected F0 value for a given pre-voicing duration based on the predictions of the linear model. No vowel-cut back was implemented for pre-voiced tokens.

All the synthesised stimuli were subsequently annotated for pre-voicing, VOT, vowel duration and F0 at the first 5ms from vowel onset. This F0 measurement was made in order to align the data with the production database that we use for ideal observer analysis. Each item's F0 in relation to VOT is plotted in figure X.

```
##
## Call:
## lm(formula = f0_5ms_into_vowel ~ 1 + VOT, data = d)
```

```
##
## Coefficients:
## (Intercept)          VOT
##    245.4697       0.0383
```

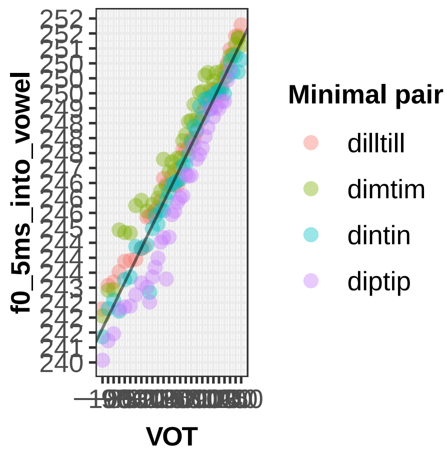

Figure 6

§3.5.1 Making exposure conditions

§4 Web-based experiment design procedure

§4.1 Norming experiment: Listener's expectations prior to informative exposure

The norming experiment 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.

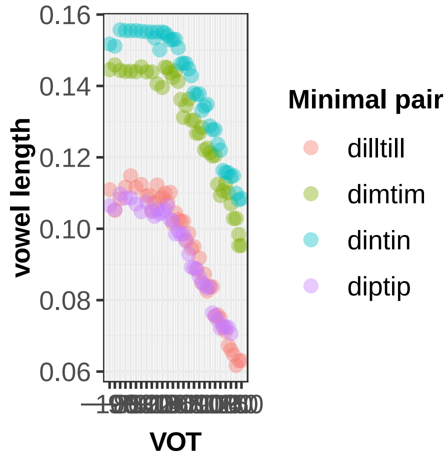

Figure 7

The primary goal of norming was methodological. We used the norming experiment to test basic assumptions about the paradigm and stimuli we employ in 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 the main experiment.* 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 the norming experiment, 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 the main.

§4.2 Methods

§4.2.1 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.

§4.2.2 Materials

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.

§4.2.3 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 hear 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 **??**. 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, 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.

§4.2.4 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.

§4.2.5 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 et al., 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 et al., 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.

§4.2.6 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).

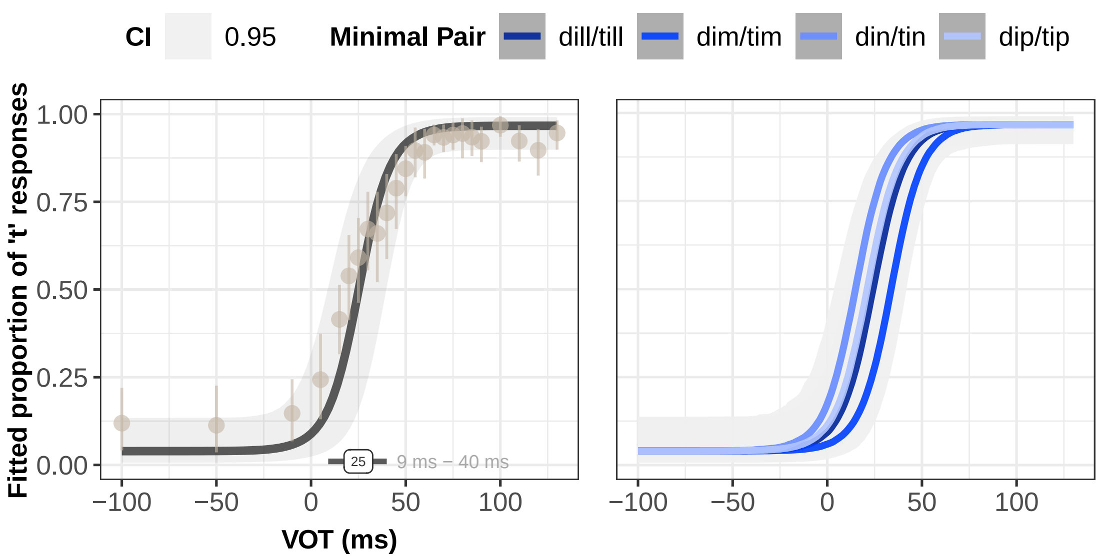

Figure 8. Categorisation functions and points of subjective equality (PSE) derived from the Bayesian mixed-effects psychometric model fit to listeners' responses in Experiment 1. The categorization functions include lapse rates and biases. The PSEs correct for lapse rates and lapse biases (i.e., they are the PSEs of the perceptual component of the psychometric model).[2] **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.

The lapse rate was estimated to be on the slightly larger side, but within the expected 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 8 (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 8 (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 8 (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.

§4.2.7 Regression analysis - model selection

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

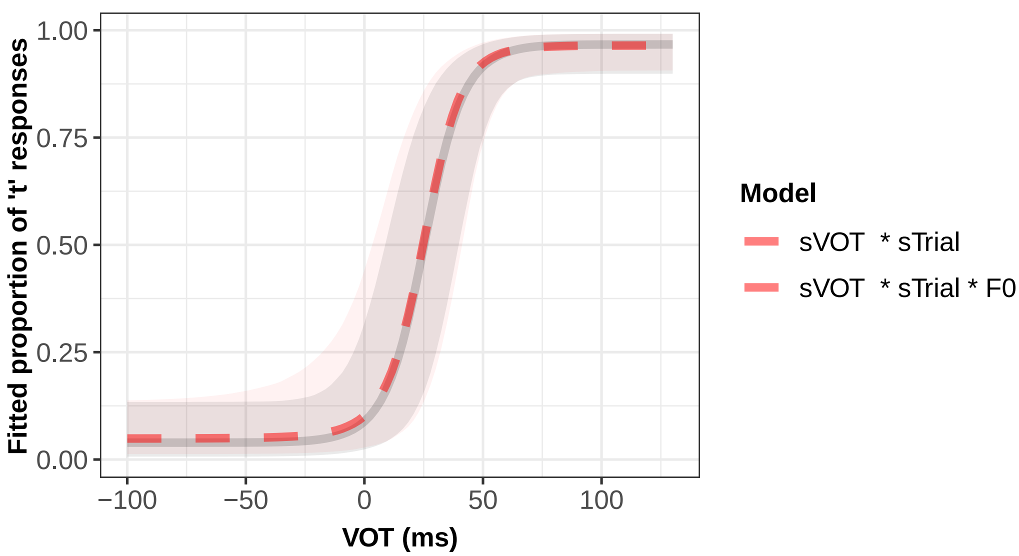

Figure 9. Expected effect of VOT interacting with trial on categorisation from model: 1 + (sVOT + sFO) * sTrial shown as red dashed line with pink shaded CI. Grey line and shaded area represents effects of VOT interacting with trial from model: 1 + sVOT * sTrial

§4.3 Experiment 2

§4.3.1 Exclusions analysis

- reaction time plots
- catch trial performance plots

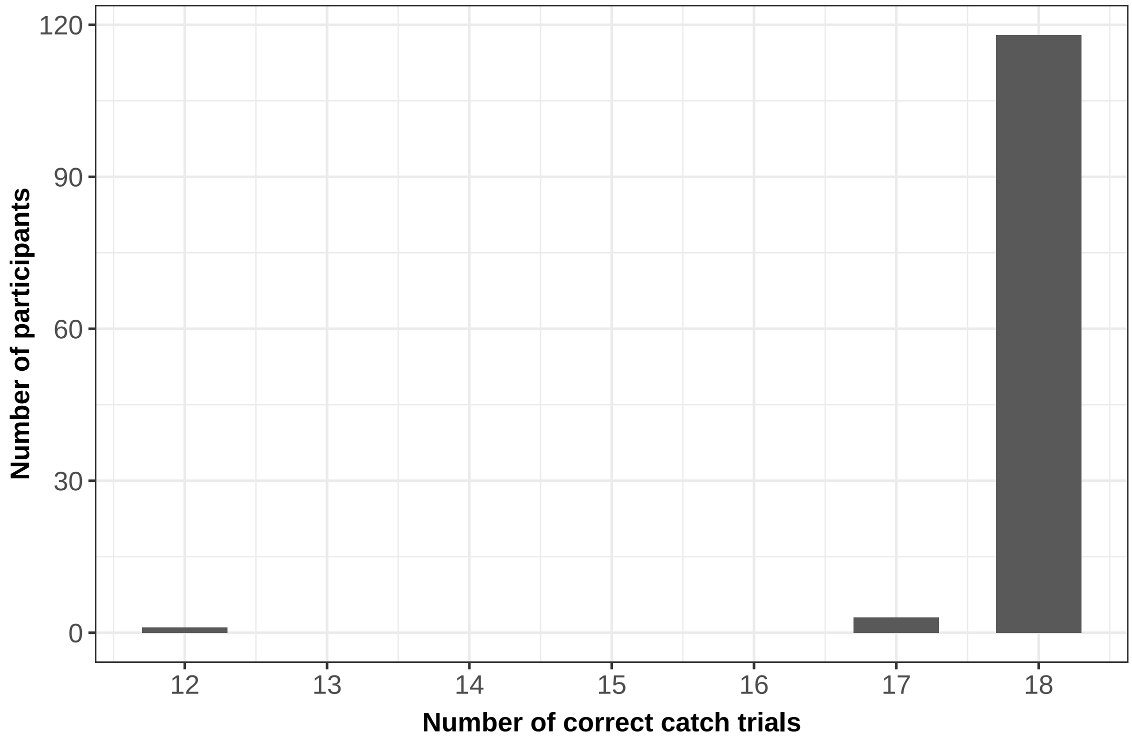

Figure 10

-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.

## 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.
```

§4.4 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

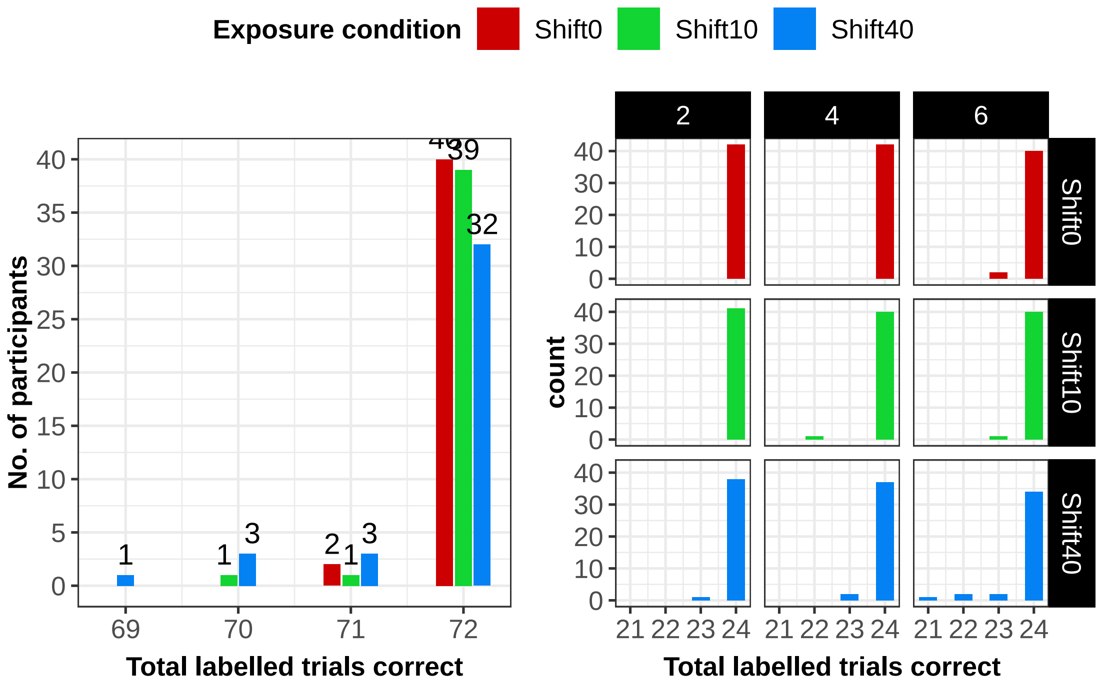

Figure 11

to derive the IOs' posterior probability of categorising the test stimuli as "t". This is defined as 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.

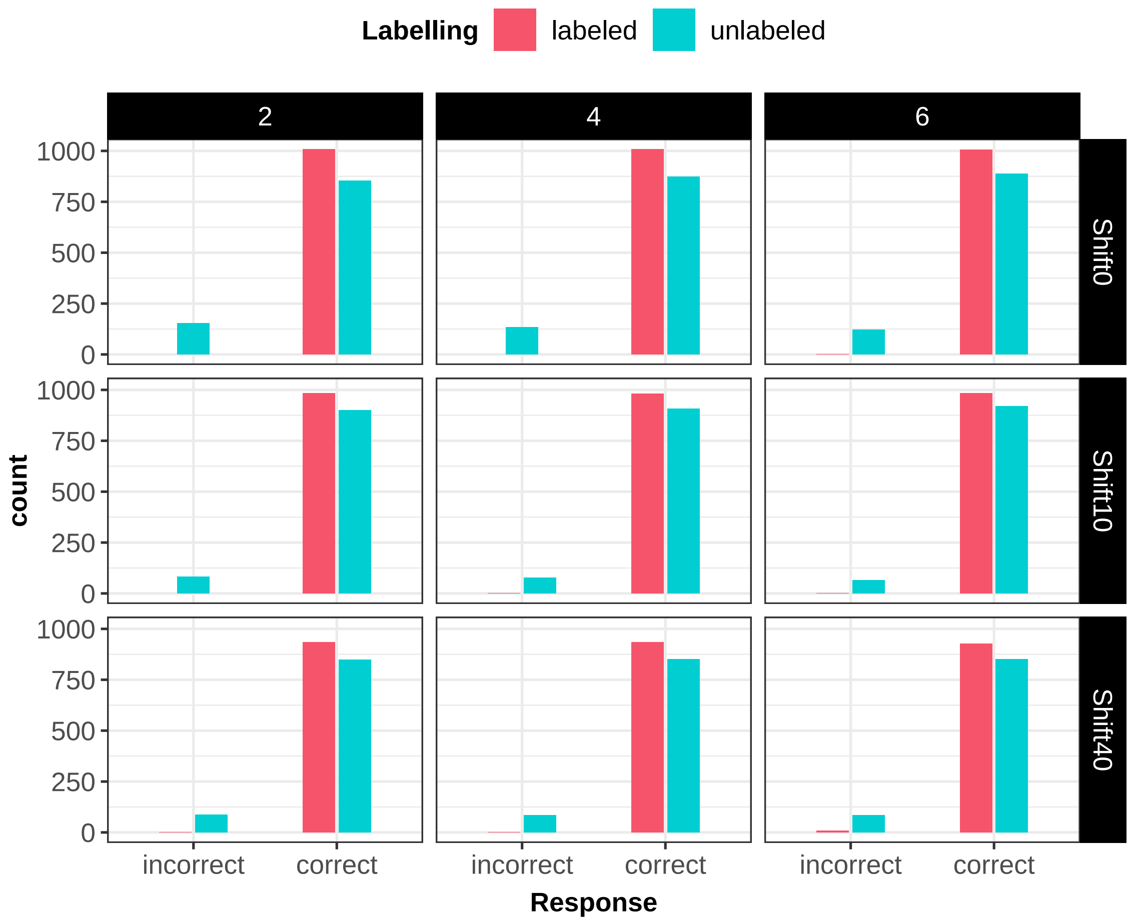

Figure 12

§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
```

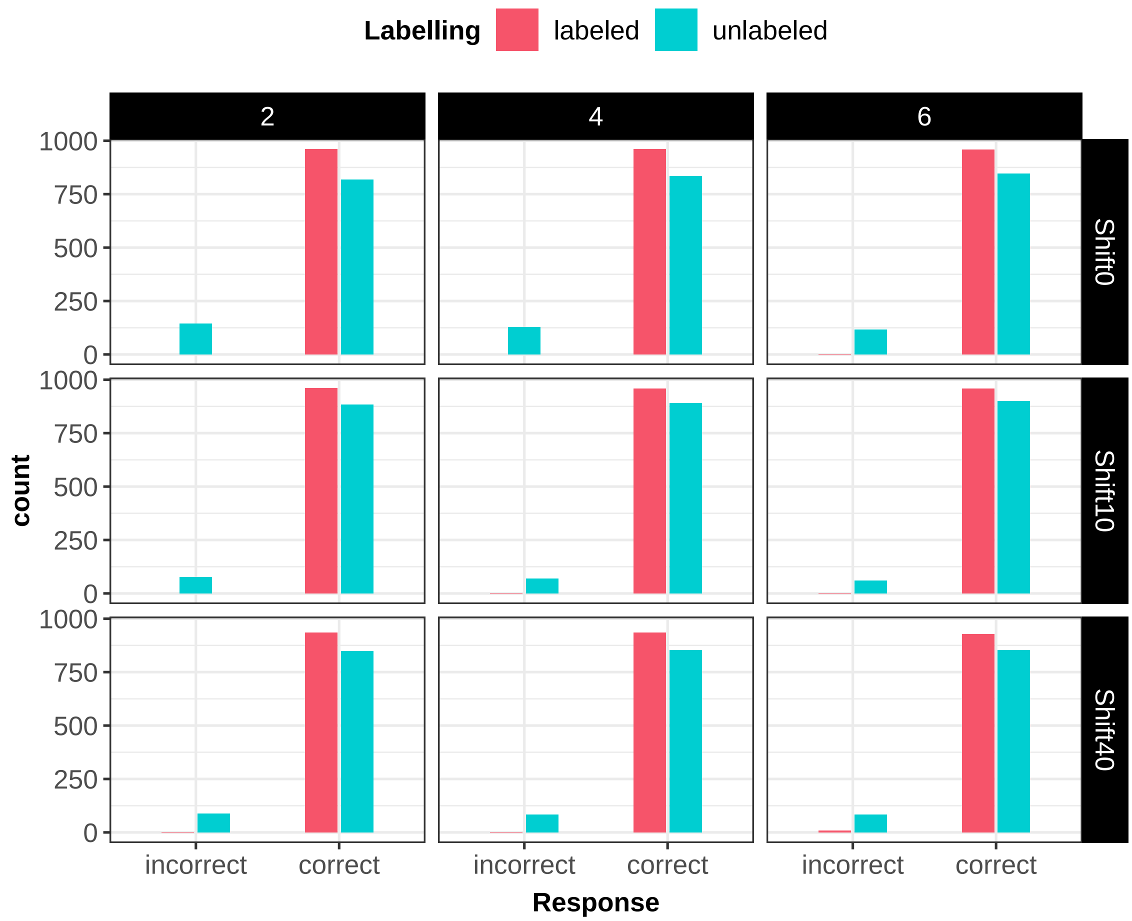

Figure 13

```
## tz         America/New_York
## date       2023-05-24
## 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)
```

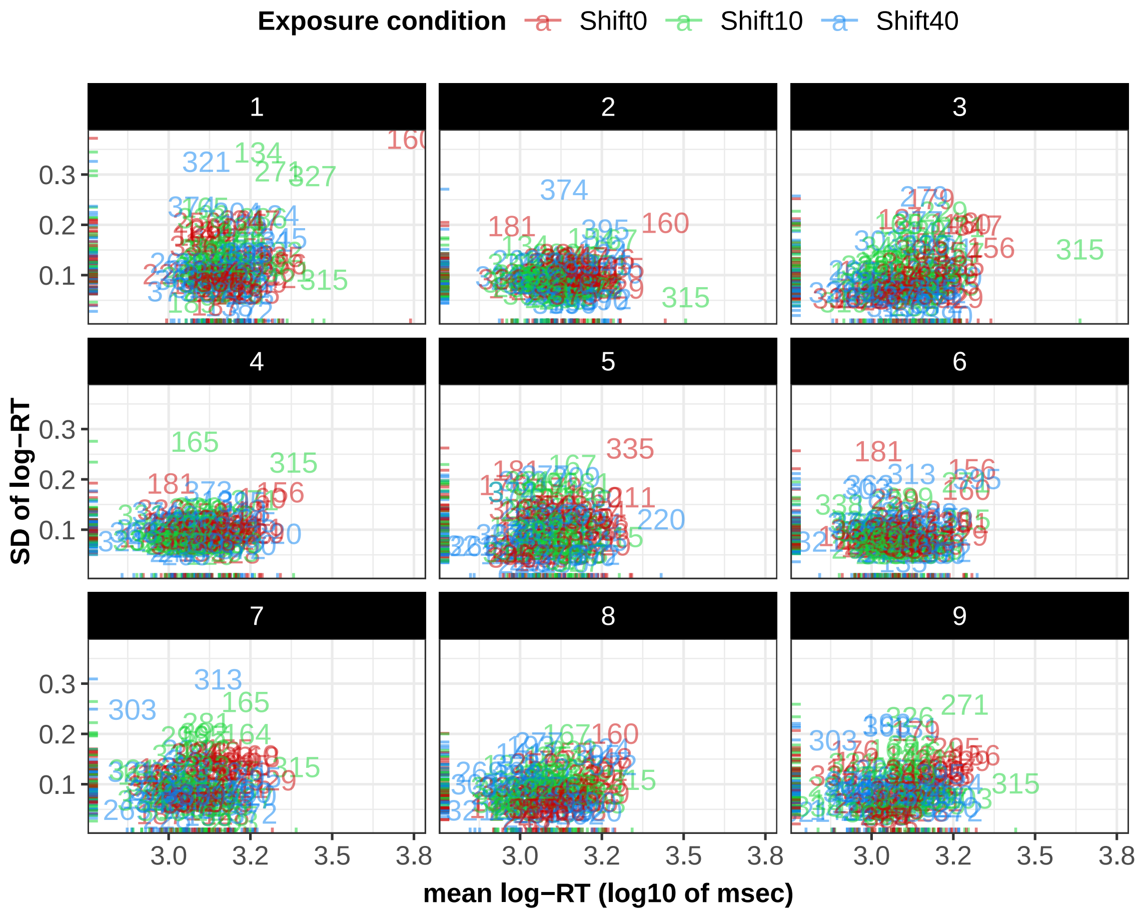

Figure 14

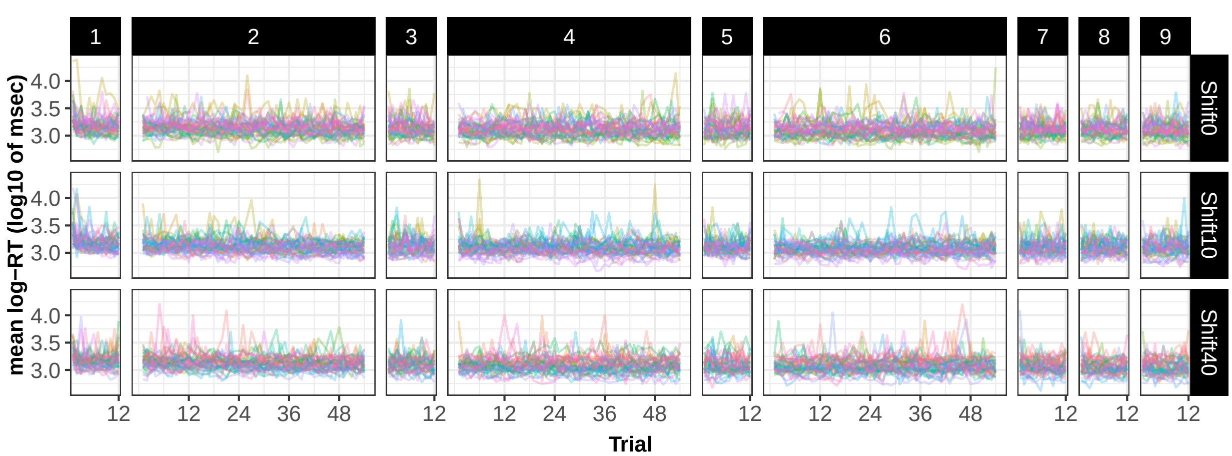

Figure 15

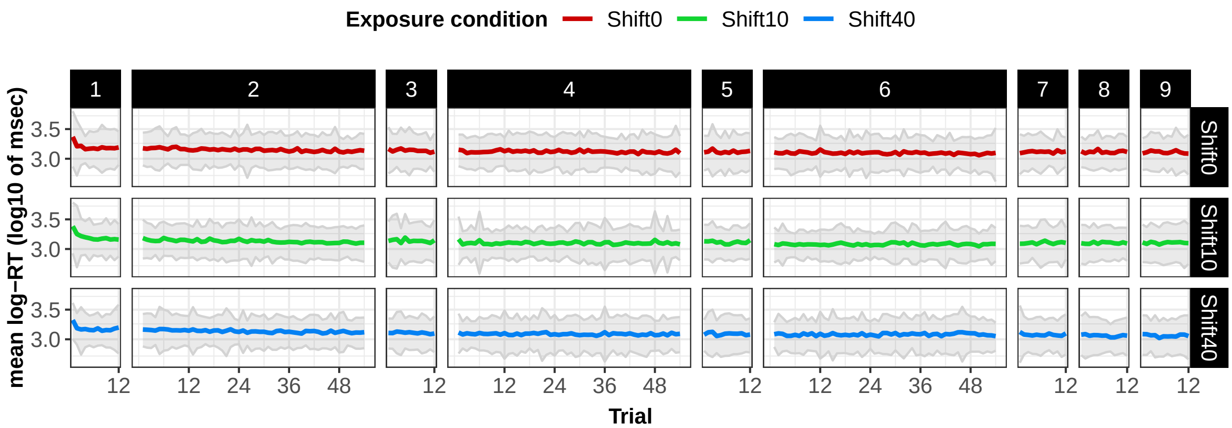

Figure 16

```
## 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       0.13.1    2023-04-07 [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         0.34      2023-05-09 [1] CRAN (R 4.1.3)
## 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           * 2.19.0    2023-03-14 [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           1.0.8     2023-05-01 [1] CRAN (R 4.1.2)
## callr            3.7.3     2022-11-02 [1] CRAN (R 4.1.2)
## car              3.1-2     2023-03-30 [1] CRAN (R 4.1.2)
## carData          3.0-5     2022-01-06 [1] CRAN (R 4.1.2)
## checkmate        2.2.0     2023-04-27 [1] CRAN (R 4.1.2)
## class            7.3-22    2023-05-03 [1] CRAN (R 4.1.2)
## classInt         0.4-9     2023-02-28 [1] CRAN (R 4.1.2)
## cli              3.6.1     2023-03-23 [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          0.2-19   2023-02-01 [1] CRAN (R 4.1.2)
## 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             * 5.0.0    2023-01-12 [1] CRAN (R 4.1.2)
## data.table         1.14.8   2023-02-17 [1] CRAN (R 4.1.2)
## datawizard         0.7.1    2023-04-03 [1] CRAN (R 4.1.2)
## DBI                1.1.3    2022-06-18 [1] CRAN (R 4.1.2)
## 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     0.3.2    2023-03-22 [1] CRAN (R 4.1.2)
## dplyr            * 1.1.2    2023-04-20 [1] CRAN (R 4.1.2)
## DT                 0.28     2023-05-18 [1] CRAN (R 4.1.3)
## 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         0.8.3    2023-01-28 [1] CRAN (R 4.1.2)
## ellipse            0.4.5    2023-04-05 [1] CRAN (R 4.1.2)
## ellipsis           0.3.2    2021-04-29 [1] CRAN (R 4.1.0)
## emmeans            1.8.6    2023-05-11 [1] CRAN (R 4.1.2)
## estimability       1.4.1    2022-08-05 [1] CRAN (R 4.1.2)
## evaluate           0.21     2023-05-05 [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          0.8-84   2022-12-06 [1] CRAN (R 4.1.2)
## Formula          1.2-5    2023-02-24 [1] CRAN (R 4.1.3)
## fs               1.6.2    2023-04-25 [1] CRAN (R 4.1.2)
## furrr            0.3.1    2022-08-15 [1] CRAN (R 4.1.2)
## future           1.32.0   2023-03-07 [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           3.3.0    2023-05-13 [1] CRAN (R 4.1.3)
## ggeffects        1.2.2    2023-05-04 [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        * 3.4.2    2023-04-03 [1] CRAN (R 4.1.2)
## ggpubr           0.6.0    2023-02-10 [1] CRAN (R 4.1.2)
## ggrepel          0.9.3    2023-02-03 [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           1.12.0   2023-05-19 [1] CRAN (R 4.1.3)
## 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)
## gridExtra        2.3      2017-09-09 [1] CRAN (R 4.1.0)
## gtable           0.3.3    2023-03-21 [1] CRAN (R 4.1.2)
## gtools           3.9.4    2022-11-27 [1] CRAN (R 4.1.2)
## Hmisc            5.1-0    2023-05-08 [1] CRAN (R 4.1.2)
## hms              1.1.3    2023-03-21 [1] CRAN (R 4.1.2)
## htmlTable        2.4.1    2022-07-07 [1] CRAN (R 4.1.2)
## htmltools        0.5.5    2023-03-23 [1] CRAN (R 4.1.2)
## htmlwidgets      1.6.2    2023-03-17 [1] CRAN (R 4.1.2)
```

```
## httpuv            1.6.11   2023-05-11 [1] CRAN (R 4.1.3)
## httr              1.4.6    2023-05-08 [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           0.19.2   2023-05-23 [1] CRAN (R 4.1.3)
## isoband           0.2.7    2022-12-20 [1] CRAN (R 4.1.2)
## iterators         1.0.14   2022-02-05 [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        2.23-21  2023-05-03 [1] CRAN (R 4.1.2)
## 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             1.3.1    2023-05-02 [1] CRAN (R 4.1.2)
## latexdiffr      * 0.1.0    2021-05-03 [1] CRAN (R 4.1.0)
## lattice           0.21-8   2023-04-05 [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           0.9.0    2022-12-16 [1] CRAN (R 4.1.2)
## lme4            * 1.1-33   2023-04-25 [1] CRAN (R 4.1.2)
## lmerTest          3.1-3    2020-10-23 [1] CRAN (R 4.1.0)
## loo               2.6.0    2023-03-31 [1] CRAN (R 4.1.2)
## lpSolve           5.6.18   2023-02-01 [1] CRAN (R 4.1.2)
## lubridate       * 1.9.2    2023-02-10 [1] CRAN (R 4.1.2)
## magick          * 2.7.4    2023-03-09 [1] CRAN (R 4.1.2)
## magrittr        * 2.0.3    2022-03-30 [1] CRAN (R 4.1.2)
## markdown          1.7      2023-05-16 [1] CRAN (R 4.1.3)
## MASS            * 7.3-60   2023-05-04 [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           0.1.11     2023-03-22 [1] CRAN (R 4.1.2)
## multcomp         1.4-23     2023-03-09 [1] CRAN (R 4.1.2)
## munsell          0.5.0      2018-06-12 [1] CRAN (R 4.1.0)
## MVBeliefUpdatr * 0.0.1.0002 2023-05-19 [1] Github (hlplab/MVBeliefUpdatr@fae8746)
## mvtnorm          1.1-3      2021-10-08 [1] CRAN (R 4.1.0)
## nlme             3.1-162    2023-01-31 [1] CRAN (R 4.1.2)
## nloptr           2.0.3      2022-05-26 [1] CRAN (R 4.1.2)
## nnet             7.3-19     2023-05-03 [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 2023-05-09 [1] Github (crsh/papaja@1c488f7)
## parallelly       1.35.0     2023-03-23 [1] CRAN (R 4.1.2)
## parameters       0.21.0     2023-04-19 [1] CRAN (R 4.1.2)
## patchwork      * 1.1.2      2022-08-19 [1] CRAN (R 4.1.2)
## performance      0.10.3     2023-04-07 [1] CRAN (R 4.1.2)
## phonR          * 1.0-7      2016-08-25 [1] CRAN (R 4.1.0)
## pillar           1.9.0      2023-03-22 [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      * 1.4.1      2023-03-14 [1] CRAN (R 4.1.2)
```

```
## prettyunits       1.1.1      2020-01-24 [1] CRAN (R 4.1.0)
## processx          3.8.1      2023-04-18 [1] CRAN (R 4.1.2)
## profvis           0.3.8      2023-05-02 [1] CRAN (R 4.1.2)
## 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                1.7.5      2023-04-18 [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      5.1.7      2023-02-27 [1] CRAN (R 4.1.2)
## Rdpack            2.4        2022-07-20 [1] CRAN (R 4.1.2)
## readr           * 2.1.4      2023-02-10 [1] CRAN (R 4.1.2)
## remotes           2.4.2      2021-11-30 [1] CRAN (R 4.1.0)
## reshape2          1.4.4      2020-04-09 [1] CRAN (R 4.1.0)
## rlang           * 1.1.1      2023-04-28 [1] CRAN (R 4.1.2)
## rmarkdown         2.21       2023-03-26 [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        2.3.1      2023-03-30 [1] CRAN (R 4.1.2)
## rstatix           0.7.2      2023-02-01 [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                1.0-12     2023-03-19 [1] CRAN (R 4.1.2)
```

```
## shiny            1.7.4    2022-12-15 [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)
## sjlabelled       1.2.0    2022-04-10 [1] CRAN (R 4.1.2)
## sjmisc           2.8.9    2021-12-03 [1] CRAN (R 4.1.0)
## sjPlot         * 2.8.14   2023-04-02 [1] CRAN (R 4.1.2)
## sjstats          0.18.2   2022-11-19 [1] CRAN (R 4.1.2)
## StanHeaders      2.26.25  2023-05-17 [1] CRAN (R 4.1.3)
## 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         3.5-5    2023-03-12 [1] CRAN (R 4.1.2)
## svglite          2.1.1    2023-01-10 [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)
## terra          * 1.7-29   2023-04-22 [1] CRAN (R 4.1.2)
## TH.data          1.1-2    2023-04-17 [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      * 3.0.4    2023-03-14 [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      * 2.0.0    2023-02-22 [1] CRAN (R 4.1.2)
## timechange       0.2.0    2023-01-11 [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             0.4.0    2023-05-12 [1] CRAN (R 4.1.3)
```

```
## units             0.8-2    2023-04-27 [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             0.6.2    2023-04-19 [1] CRAN (R 4.1.2)
## viridis           0.6.3    2023-05-03 [1] CRAN (R 4.1.2)
## viridisLite       0.4.2    2023-05-02 [1] CRAN (R 4.1.2)
## vroom             1.6.3    2023-04-28 [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              0.39     2023-04-20 [1] CRAN (R 4.1.2)
## xml2              1.3.4    2023-04-27 [1] CRAN (R 4.1.2)
## xtable            1.8-4    2019-04-21 [1] CRAN (R 4.1.0)
## xts               0.13.1   2023-04-16 [1] CRAN (R 4.1.2)
## yaml              2.3.7    2023-01-23 [1] CRAN (R 4.1.2)
## zoo               1.8-12   2023-04-13 [1] CRAN (R 4.1.2)
##
## [1] /Library/Frameworks/R.framework/Versions/4.1/Resources/library
##
## ------------------------------------------------------------------------------------------
```