Perceptual learning of multiple talkers: Determinants, characteristics, and limitations

Shawn N. Cummings and Rachel M. Theodore

Department of Speech, Language, and Hearing Sciences
University of Connecticut
2 Alethia Drive, Unit 1085
Storrs, CT 06269-1085
United States

Connecticut Institute for the Brain and Cognitive Sciences
University of Connecticut
337 Mansfield Road, Unit 1272
Storrs, CT 06269-1272
United States

Corresponding author:
Rachel M. Theodore
rachel.theodore@uconn.edu
(860) 486-3477

Abstract

Research suggests that listeners simultaneously update talker-specific generative models to reflect structured phonetic variation. Because past investigations exposed listeners to talkers of different genders, it is unknown whether adaptation is talker-specific or rather linked to a broader sociophonetic class. Here we test determinants of listeners' ability to update and apply talker-specific models for speech perception. In six experiments (*n* = 480), listeners were first exposed to the speech of two talkers who produced ambiguous fricative energy. The talkers' speech was interleaved during exposure, and lexical context differentially biased interpretation of the ambiguity as either /s/ or /ʃ/ for each talker. At test, listeners categorized tokens from *ashi – asi* continua, one for each talker. Across conditions and experiments, we manipulated exposure quantity, talker gender, blocked versus interleaved talker structure at test, and the degree to which fricative acoustics differed between talkers. When test was blocked by talker, learning was observed for different but not same gender talkers. When talkers were interleaved at test, learning was observed for both different and same gender talkers, which was attenuated when fricative acoustics were constant across talkers. There was no strong evidence to suggest that adaptation to multiple talkers required increased quantity of exposure beyond that required to adapt to a single talker. These results suggest that perceptual learning for speech is achieved via a mechanism that represents a context-dependent, cumulative integration of experience with speech input and identity critical constraints on listeners' ability to dynamically apply multiple generative models in mixed talker listening environments.

Introduction

Robust speech perception requires listeners to resolve an extensive computational hurdle; namely, there is no one-to-one relationship between the acoustics patterns in speech input and a speaker's intended linguistic message (Liberman et al., 1967). Individual talkers differ in how they instantiate speech sounds, and thus *who* is speaking serves as a primary source of variability that contributes to the lack of invariance between speech acoustics and speech segments. Talker differences in phonetic properties of speech can reflect physiological aspects of the talker (Fant, 1973; Peterson & Barney, 1952), sociophonetic characteristics (Byrd, 1992; Johnson & Beckman, 1997; Klatt, 1986), and even idiosyncratic differences in pronunciations habits (Allen et al., 2003; Chodroff & Wilson, 2017; Hillenbrand et al., 1995; Johnson & Beckman, 1997; Theodore et al., 2009). Because of these talker differences in speech production, one talker's production of a given speech category (e.g., the /s/ in *sun*) may be acoustically identical to a different talker's production of a different speech category (e.g., the /ʃ/ in *shun*) (Newman, Clouse, & Burnham, 2001). A theoretical account of how listeners map acoustics to meaning given extensive variability in speech input remains an unsolved challenge in the domain of speech perception (Saltzman, et al., 2021; Liberman, et al., 1957).

The lack of invariance in speech input requires that perceptual systems also be not invariant. Classification models for speech based on naïve invariance cannot achieve similar accuracy to human listeners, even when using as many as 24 unique and informative acoustic cues towards phoneme identity (McMurray & Jongman, 2011). Rather, perception must be able to *adapt* in order to accommodate contextual variation in the input, such as a particular talker, a group of talkers who share social and physiological characteristics, or a particular environment wherein speech is predictably and systematically altered (Kleinschmidt & Jaeger, 2015). Indeed, human listeners adapt constantly to the input around them (Bradlow & Bent, 2008; Clopper & Pisoni, 2004; Drouin et al., 2016; Giovannone & Theodore, 2021; Kraljic & Samuel, 2005; Norris et al., 2003; Sidaras et al., 2009; Tarabeih-Ghanayim et al., 2020; Theodore et al., 2015, 2019; Theodore & Miller, 2010; Theodore & Monto, 2019; Tzeng et al., 2021; Weatherholtz & Jaeger, 2016). In some cases, supervisory signals (e.g., lexical guidance, audiovisual cues, orthographic cues) help to disambiguate otherwise unclear acoustics, and listeners can leverage this guidance to inform interpretation of subsequent input (Bertelson et al., 2003; Drouin & Theodore, 2018; Keetels et al., 2016; Norris et al., 2003; Samuel & Kraljic, 2009; Tzeng et al., 2021). In other cases, listeners adjust expectations without explicit supervision via sensitivity to underlying statistical regularities in speech (e.g., Idemaru & Holt, 2014; Liu & Holt, 2015; McMurray et al., 2009; Theodore et al., 2019; Theodore & Monto, 2019).

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

Rather than representing speech sound knowledge as a single generative model for a given speech sound, the ideal adapter framework posits that listeners maintain multiple models, each specific to a given situation, such as an individual talker (Kleinschmidt & Jaeger, 2015). This is not without cost. Compared to a system that only tracks a single model, maintaining talker-specific cue distributions requires a finer grain of statistical information to be represented in memory. Additionally, increasing specificity of model representation tracks inversely with the amount of input informative to that model; that is, representing experience by talker-specific models necessarily leads to fewer observations in each talker's model compared to a model that aggregates observations across talkers. Consequently, talker-specific models may be less reliable than aggregate representations. Kleinschmidt and Jaeger (2015) acknowledge these costs and propose generalization across talkers (towards a more general, group-specific model) as a potential means to ameliorate them. They additionally identify a trade-off between the cost of maintaining multiple models and the increased specificity of speech perception that they may provide. As between-talker variation increases, the benefits of maintaining talker-specific models may outweigh or justify any cognitive or perceptual costs of maintaining multiple models. Conversely, generalizing to a group-level model may be warranted given more acoustically similar talkers.

Another factor relevant to the trade-off between the cost of maintaining multiple models and the increased specificity of speech perception is the frequency that a listener is required to dynamically retrieve different models. In a multi-talker listening environment, a system reliant on talker-specific models may require listeners to switch the model they use (and update) on potentially every utterance. A more general architecture, such as maintaining and retrieving a group-level model, may allow listeners to maintain a single general model for all speech within a given listening situation. The ideal adapter framework remains agnostic as to whether dynamically switching between models in online speech perception incurs any cost to the listener. Also unspecified are whether there may be limits on the number of models that can be simultaneously maintained or the speed at which models may be dynamically retrieved. However, previous research examining adaptation to multiple talkers has posited a cost when listeners must choose which generative model to use (Luthra et al., 2021).

In this context, the goal of the current work is to test the broad hypothesis that perceptual learning for speech reflects listeners' ability to maintain *and* retrieve talker-specific models. Here we use a modified version of the widely influential lexically guided perceptual learning paradigm to examine talker-specificity of perceptual learning given a mixed talker listening environment (Drouin et al., 2016; Drouin & Theodore, 2018; Eisner & McQueen, 2005; Liu & Jaeger, 2018; Myers & Mesite, 2014; Norris et al., 2003; Reinisch & Holt, 2014; Samuel & Kraljic, 2009; Tzeng et al., 2021). Below we provide an overview of lexically guided perceptual learning for speech, summarize key findings in the extant literature regarding talker-specificity of lexically guided perceptual learning, and then introduce the current experiments.

Lexically guided perceptual learning

Evidence of lexically guided perceptual learning has rich empirical support and recent work has established that learning in this paradigm follows predictions made by the ideal adapter framework (Liu & Jaeger, 2018; Luthra et al., 2021; Saltzman & Myers, 2021; Tzeng et al., 2021). The standard paradigm includes two phases, exposure and test. During exposure, listeners hear speech from a single talker. Acoustic energy of canonical sounds is replaced with acoustic energy that is perceptually ambiguous between two speech sounds. For example, canonical productions of /s/ or /ʃ/ are replaced by spectral energy perceptually ambiguous between /s/ and /ʃ/. A supervisory signal is provided, such that the listener is aware of the intended category of the ambiguous sound. Most commonly, this supervisory signal is lexical – achieved by embedding the ambiguous sound in items that form real words *only* if they are interpreted as one category. For example, replacing the canonical /s/ in *personal* with ambiguous spectral energy allows surrounding phonetic context to guide listeners to interpret the energy as /s/ because *personal* is an English word and *pershonal* is not. Likewise, replacing the canonical /ʃ/ in *publisher* with ambiguous spectral energy allows phonetic context to guide listeners to interpret the energy as /ʃ/ because *publisher* is a real word but *publiser* is not. Lexical context biases interpretation of the ambiguous sound towards the category that supports lexical access (Ganong, 1980). In the standard paradigm, lexical bias is manipulated between subjects (e.g., one group of listeners is biased to perceive the ambiguity as /s/ and a different group is biased to perceive the ambiguity as /ʃ/), which allows listeners in each bias group to differentially build a generative model for the exposure talker. The most common task used during exposure is a lexical decision task, though lexically guided perceptual learning is not contingent on explicit lexical decision (Drouin & Theodore, 2018; Jesse, 2021; Keetels et al., 2016; Luthra et al., 2021; McQueen et al., 2006; Samuel, 2016; van Linden & Vroomen, 2007). Following exposure, listeners complete a phonetic identification test phase that is identical between bias groups, with stimuli drawn from a continuum that spans the two categories manipulated during exposure (e.g., tokens from an *ashi* to *asi* continuum). Evidence of learning manifests as a difference in performance at test, indicating that listeners modified the mapping between acoustics and meaning in line with lexical bias during exposure (e.g., more *asi* responses for listeners in the /s/-bias compared to the /ʃ/-bias exposure group).

Previous research examining talker-specificity of lexically guided perceptual learning

Findings from the perceptual learning domain have provided myriad contributions to a theoretical understanding of dynamic adaptation in speech perception, including an understanding of situations that promote talker-specific learning versus generalization across talkers. Two factors of interest for the current work include the role of between-talker acoustic similarity and the flexibility of model adaptation and retrieval. The ideal adapter framework posits that talker-specificity is warranted when "variation *between* talkers is so large that the differences in the cue distributions of phonetic categories *within* each talker are obscured" (Kleinschmidt & Jaeger, 2015). Generalization across talkers is thus predicted to occur when between-talker variability is relatively minimal, whereases specificity of learning is predicted to occur when between-talker variability is relatively maximal. However, support for this prediction is limited. For example, given exposure to a female talker, Eisner and McQueen (2005) observed generalization to a novel female *and* a novel male talker when the novel talker was cued by the vowel (and not fricative) portions of the test continuum. Tamminga et al. (2020) exposed listeners to one of four female talkers, and then tested generalization to either a novel female talker or a novel male talker. Robust generalization was found to the novel female talker, but not to the novel male talker. Whether generalization was driven by acoustic similarity between the (female) exposure talker and the female test talker, or rather by the shared indexical trait of gender, is an open question.

Further evidence of generalization across talkers comes from Kraljic and Samuel (2005), who exposed listeners to *either* a male or female talker and then tested learning for *both* talkers. Exposure to the male talker resulted in learning for the male talker, which did not generalize to the female talker. In contrast, however, exposure to the female talker *did* result in learning that generalized to the male talker. This unique pattern of results was explained by post hoc acoustic analyses that showed a low degree of acoustic similarity between the male exposure tokens and the female test tokens and a high degree of acoustic similarity between the female exposure tokens and the male test tokens. These results highlight an important role of acoustic similarity between talkers for perceptual learning; talkers with similar acoustic characteristics may constitute a single situation in the ideal adapter framework and thus may share a single generative model linked to acoustic similarity rather than talker or gender.

Though studies that examine generalization of learning given exposure to a *single* talker, reviewed above, yield mixed evidence in support of talker-specificity of learning, other studies that examine learning given exposure to *two talkers* are consistent with talker-specific learning. Kraljic and Samuel (2007) modified the standard lexically guided learning paradigm to provide exposure to two talkers that differed in gender, with lexical information used to bias perception of an ambiguous fricative in opposite directions such that the ambiguous fricative mapped to /s/ for one talker and /ʃ/ for the other talker. Exposure and test trials were both blocked by talker. Learning was observed, suggesting that listeners adapted separate models for each talker. This finding was replicated by Luthra et al. (2021) and extended to show that listeners could maintain talker-specific generative models when input from each talker was *interleaved* during exposure. However, learning given that interleaved exposure was only observed when the exposure dose was doubled (32 critical exposures) compared to the dose required to invoke learning given blocked talker exposure (16 critical exposures). That is, Luthra et al. (2021) posit that there is a processing cost associated with switching between talker-specific models given interleaved talker exposure that results in an adaptation process that requires twice as much exposure to learn a talker's cue distributions compared to exposure that is blocked by talker. Luthra et al. (2021) interpreted these results as being consistent with domain-general learning theories, citing findings within motor-skill learning where high trial-by-trial variability facilitates high contextual interference, which in turn leads to poorer performance (Magill & Hall, 1990; Shea & Morgan, 1979).

Though these studies have provided critical insight towards a theoretical understanding of talker-specificity and acoustic similarity for perceptual learning, they do not completely explicate the role of model adaptation and retrieval because test (and, in most cases, exposure too) were always blocked by talker. That is, blocking test by talker does not require listeners to *dynamically* retrieve talker-specific generative models as would be required given trial-by-trial talker variability at test. A meaningful distinction should be made here between choosing which model to update (given supervisory lexical signals during exposure) and choosing which model to use for categorization (during test, in which no supervisory lexical information is available). Interleaving talkers' input during exposure is a critical first step towards testing the hypothesis that listeners adapt talker-specific generative models; however, to test whether listeners dynamically retrieve a talker's specific model *in categorization* of speech, trial-by-trial talker variability must be present when listeners are performing speech categorization. In addition, interleaving talkers' speech during exposure *and* test promotes ecological validity in simulating a mixed talker listening environment and removes a memory-based confound in the blocked test design in which time between exposure and test differs between talkers.

Introducing the current experiments

The goal of the current work is to test the hypothesis that perceptual learning for speech reflects listeners' ability to maintain *and* retrieve talker-specific models. Six experiments were conducted; a summary of the manipulations across experiments is shown in Table 1. Listeners in each experiment were exposed to the speech of two talkers during a talker identification exposure phase in which words produced by each talker were randomly interleaved in time. The phonetic input for each of the two talkers was manipulated such that lexical information was used to bias perception of ambiguous spectral energy as /s/ for one talker and as /ʃ/ for the other talker. That is, lexical information was used to elicit perceptual recalibration in opposite directions for the two talkers heard during exposure. To provide a replication of the primary finding Luthra et al. (2021) – that simultaneous adaptation to multiple talkers requires increased exposure – each experiment manipulated exposure dose to include either a standard dose of 20 critical productions for each talker, or a doubling of that dose, yielding 40 critical productions for each talker. We defined 20 tokens as our standard dose because this dose reflects the quantity of exposure most often provided in previous investigations of lexically guided perceptual learning. Following exposure, all listeners completed a test phase in which learning was assessed for both talkers using a phonetic identification task.

In experiment 1, the two exposure talkers differed in gender and, critically, the test phase was blocked by talker. Accordingly, experiment 1 provides a direct replication of Luthra et al. (2021). In experiment 2, speech from the two (different gender) talkers is interleaved at exposure *and* at test; thus, listeners are required to make frequent, dynamic changes to the generative model they use at test in experiment 2 compared to experiment 1. If adaptation reflects dynamic retrieval of separate generative models, then learning will be observed even in the face of trial-by-trial talker variability at test. Attenuated or absent learning in experiment 2 compared to experiment 1 would severely constrain the claim that listeners use unique generative models to guide online perception. Experiment 3 probes the extent to which perception of acoustics is conditioned on talker by holding the critical fricative acoustics constant across the two (different gender) talkers. Qualitative reasoning in the lexically guided perceptual learning domain often assumes this level of control to guide conclusions that perception of speech acoustics has been conditioned on talker identity; here we examine this hypothesis directly. Learning in experiment 3 is expected if, and only if, listeners maintain separate generative models.

Experiments 4 – 6 parallel experiments 1 – 3 with one key exception; instead of being exposed to two talkers that differ in gender, listeners heard speech from two women during exposure. Accordingly, the two talkers' voices are more similar in experiments 4 – 6 compared to experiments 1 – 3. Past research suggests that acoustic similarity may be a driving force for generalization across talkers (e.g., Kraljic & Samuel, 2005) and the ideal adapter framework under consideration here specifically posits that adaptation may be linked to gender-specific instead of talker-specific models. To our knowledge, all work to date, if involving exposure to more than one talker, has included talkers of different genders, and thus evidence in support of talker-specificity for lexically guided perceptual learning has been confounded with gender (Kraljic & Samuel, 2007; Luthra et al., 2021). Thus, experiments 4 – 6 provide a strict test of talker-specific adaptation by assessing whether learning occurs for two talkers of the same gender, which is predicted to occur if learning is talker-specific. In contrast, evidence of learning in experiments 1 – 3 but not experiments 4 – 6 would suggest that listeners learn and retrieve gender- but not talker-specific generative models.

Within each of the six experiments, the quantity of exposure was also manipulated to reflect a standard exposure dose (i.e., 20 critical productions per talker) or twice the standard exposure dose (i.e., 40 critical productions per talker). If perceptual learning for multiple talkers requires increased exposure, as suggested by Luthra et al. (2021), then any observed patterns of learning will be limited to the extended dose exposure conditions.

Experiment 1

Methods

Participants. Participants (*n* = 80) were recruited from the Prolific participant pool (Palan & Schitter, 2018) according to the following criteria: between 18 – 35 years of age, monolingual English speaker born in and currently residing in the US, no history of language related disorders, Prolific score ≥ 98, and completed ≥ 10 previous Prolific experiments. Further, we limited participation from the Prolific pool to exclude participants who had participated in any previous lexically guided perceptual learning study conducted in our laboratory and to ensure unique participants across the six experiments presented in this manuscript. The sample included 33 women and 47 men with an average age of 28 years (*SD* = 5 years). An additional three participants were tested but excluded from analyses based on preregistered exclusion criteria, which included failure to pass the headphone screen (*n* = 1) and a flat identification function at test (*n* = 2). Participants were randomly assigned to either the 1x dose (*n* = 40) or 2x dose (*n* = 40) condition.

The sample size was determined based on an a priori power analysis. Evaluating our primary hypotheses requires power to detect an effect of bias. In principle, the ability to replicate the primary finding of Luthra et al. (2021) – that perceptual learning of multiple talkers requires additional exposure – requires power to detect an interaction between bias and exposure dose.

However, as reported in Luthra et al. (2021), this interaction was not observed. Because of this, our power analyses were based on the bias effect observed in the “2x” exposure dose conditions of Luthra et al. (2021), referred to as experiments 2C and 2D. We combined data across these experiments to have the most precise estimate of the effect size for bias in the experiments where an effect of bias was reported. We executed our power analysis using the simr package (Green & MacLeod, 2016); a reproducible pipeline for the power analysis is provided in the OSF repository for this manuscript. The power analysis showed that 40 participants yields high power (97%) to detect an effect of bias of the magnitude observed in experiments 2C and 2D of Luthra et al. (2021). Accordingly, we set the sample size to 40 participants in each dose condition across all experiments. We note that an additional set of power analyses executed with the simr package (Green & MacLeod, 2016) showed that this sample size had high power (87%) to detect a bias by dose interaction of the magnitude observed in Tzeng et al. (2021), which reflected an attenuation of the lexically guided perceptual learning effect (for a single talker) given exposure to 10 versus 20 critical productions.

Stimuli. Two sets of stimuli were created, one for each of two talkers who were fictiously referred to as Joanne and Peter. Each set contained 80 exposure tokens and six test tokens. Exposure tokens were auditory recordings of 40 English words, 20 containing a single instance of /s/ and no occurrence of /ʃ/ (e.g., *rehearsal*) and the other 20 containing a single /ʃ/ and no occurrence of /s/ (e.g., *publisher*). The /s/ and /ʃ/ words follow those used in Kraljic and Samuel (2005) and were matched in mean syllable length and word frequency. Two variants of each word were created, one that contained the natural production of /s/ or /ʃ/ (the clear variant) and one in which the natural production of /s/ or /ʃ/ was replaced with a digital mixture of a natural /s/ and /ʃ/ production that was judged to be perceptually ambiguous between /s/ and /ʃ/ (the ambiguous variant). Test tokens consisted of a six-step continuum that perceptually ranged from /ɑʃi/ to /ɑsi/. The fricative portion of the test continuum was created by digitally mixing energy from natural /ʃ/ and /s/ productions in different weights to yield continuum steps that ranged from 70% /ʃ/ - 30% /s/ to 20% /ʃ/ - 80% /s/ in six equidistant steps.

The stimulus set for Joanne was a subset of tokens used in Tzeng et al. (2021), to which the reader is referred for comprehensive details on stimulus construction. The stimulus set for Peter was created by applying the Change Gender function in Praat (Boersma, 2002) to Joanne's stimuli using the parameters identified in Luthra et al. (2021), which included a formant shift ratio of 0.8, a median pitch of 100 Hz, and no change to either pitch range or duration. As in Luthra et al., these parameters were sufficient to induce a robust change in perceived gender. Figure 1 displays two acoustic measurements for Joanne and Peter's stimulus sets, including (1) fundamental frequency as a measure of the sound source, which is an important cue to talker and gender identity and (2) center of gravity, which is a measurement of spectral energy in the fricative. Fundamental frequency was measured as the mean fundamental frequency in the voiced portion of each token using the Quantify Source script in the GSU PraatTools package (Owren, 2008). Center of gravity was measured as the first spectral moment in the midpoint region of each fricative using the script developed by DiCanio (DiCanio, n.d.).

As can be viewed in Figure 1, fundamental frequency was higher for Joanne compared to Peter, consistent with the parameters used in the Change Gender function. For both talkers, center of gravity for the clear variants is lower for /ʃ/ compared to /s/, reflecting the expected relationship between center of gravity and place of articulation (e.g., Jongman et al., 2000; Newman et al., 2001). Moreover, center of gravity for the ambiguous variants falls intermediate to the clear variants, consistent with the digital signal manipulation used to create the ambiguous variants. Of note, though the Change Gender parameters do not suggest nor explicitly allow control over changes in voiceless energy, this function does introduce a scaling of center of gravity in line with the change in fundamental frequency. Specifically, center of gravity for Peter’s tokens is shifted towards lower frequencies compared to Joanne’s tokens. As we discuss in experiment 3, this scaling thus yields stimuli in which spectral characteristics for the critical /s/ and /ʃ/ energy in the tokens is not equated across talkers. All stimuli are available on the OSF repository for this manuscript.

Procedure. All experiments presented here were web-based studies hosted on the Gorilla platform (Anwyl-Irvine et al., 2020). After providing informed consent, participants completed a headphone screen, an exposure phase, and a test phase. The headphone screen used tasks reported in Woods et al. (2017) and Milne et al. (2021), which are brief, dichotic listening tasks developed to screen for headphone use in web-based experiments. Participants were given two opportunities to pass the Woods et al. screen and, if they did not pass on either of these attempts, one opportunity to pass the Milne et al. screen. If a participant failed either of the first two attempts, then they were given a reminder to put on headphones before continuing to the next attempt. As reported in the participants section of each experiment, compliance with headphone use as measured by these screens was very high, resulting in minimal attrition due to failure to pass the headphone screen.

The exposure phase consisted of either 80 trials [1x dose; 2 talkers x (20 /s/ words + 20 /ʃ/ words) x 1 repetition] or 160 trials (2x dose; 2 talkers x (20 /s/ words + 20 /ʃ/ words) x 2 repetitions) of a talker identification task, modeled after the gender identification task used in Luthra et al. (2021). Because the gender identification task was not optimal for the same gender manipulations reported in experiments 4 – 6, a talker identification task was used to hold the task constant across all experiments reported in this manuscript. Each listener heard both talkers during the exposure phase, with exposure stimuli selected to differentially bias listeners to perceive ambiguous variants as /s/ for one talker and /ʃ/ for the other talker. For the /s/-bias talker, stimuli included the ambiguous /s/ variants and the clear /ʃ/ variants. For the /ʃ/-bias talker, stimuli included the ambiguous /ʃ/ variances and the clear /s/ variants. Assignment of talker to the /s/- and /ʃ/-bias conditions was counterbalanced across listeners within each dose condition. Listeners in the 1x dose condition heard one repetition of the appropriate /s/ and /ʃ/ tokens for each talker; listeners in the 2x dose condition heard two repetitions of the appropriate /s/ and /ʃ/ tokens for each talker. In both dose conditions, exposure tokens for both talkers were presented in a different randomized order for each participant; that is, tokens from the two talkers were *interleaved* during the exposure phase. On each trial, listeners heard one exposure token and were asked to indicate the talker by clicking on one of two buttons labeled either "Joanne" or "Peter." Feedback was provided on every trial in the form of a green checkmark for correct responses and a red "X" for incorrect responses. Feedback remained on the screen for 750 ms. Trials were separated by 1000 ms, timed from the participant's response to the onset of the next auditory stimulus. In order to provide an opportunity to learn the association between the talkers' voices and their names prior to beginning the talker identification task, the 80 (or 160) exposure trials were preceded by 10 familiarization trials in which listeners heard five words produced by each talker while seeing the talker's name appear on the screen; these words were the same for each talker and did not contain any instances of either /s/ or /ʃ/.

The test phase consisted of 72 trials of a phonetic identification task, 36 trials for each talker. The 36 trials for each talker consisted of six cycles of the six steps of the test continuum; each cycle was a separate randomized order of the six continuum steps for each participant. On each trial, listeners heard one test token and were asked to indicate its identity as quickly as possible by clicking on one of two buttons labeled either “ashi” or “asi.” No feedback was provided at test, and trials were separated by 1000 ms time from the participant’s response to the onset of the next auditory stimulus. The test phase was blocked by talker. Talker order and button assignment were counterbalanced across listeners within each dose condition.

The entire procedure lasted approximately 15 minutes, and participants were paid $2.50 for their participation.

Results

Performance during the exposure phase was analyzed in terms of proportion correct talker identification, which was near ceiling across participants (*mean* = 0.99, *SD* = 0.01, *range* = 0.93 – 1.00). Performance at test was analyzed in terms of *asi* responses. To visualize test performance, we first calculated mean proportion *asi* responses for each talker and each continuum step separately for each participant. Figure 2 shows grand means calculated over by-subject means. Visual inspection suggests a robust learning effect in that proportion *asi* responses are greater for the /s/-bias talker compared to the /ʃ/-bias talker. No effect of dose is visually apparent.

To analyze these patterns statistically, trial-level responses (0 = *ashi*, 1 = *asi*) were analyzed using generalized linear mixed effects models (GLMMs) with the binomial response family as implemented in lme4 (Bates et al., 2015); the Satterthwaite approximation of degrees of freedom was used to evaluate statistical significance using lmerTest (Kuznetsova et al., 2017).[1] Following Luthra et al. (2021), separate models were constructed for each dose

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

condition. Each of these models included continuum step, talker bias, and their interaction as fixed effects. Continuum step was entered into the model in terms of percent /s/ energy in the continuum step as a scaled/centered continuous variable; talker bias (/ʃ/ = -0.5, /s/ = 0.5) was entered as a mean-centered contrast. The random effects structure consisted of random intercepts by subject and random slopes for step, talker bias, and their interaction by subject, which reflects the maximal random effects structure given the experimental design.

The full results of each model are shown in Table 2. Both models revealed a main effect of talker bias that reflected more *asi* responses for the /s/-bias talker compared to the /ʃ/-bias talker (1x: $\hat{\beta} = 2.218$, $SE = 0.501$, $z = 4.428$, $p < 0.001$; 2x: $\hat{\beta} = 1.463$, $SE = 0.609$, $z = 2.402$, $p = 0.016$), indicative of lexically guided perceptual learning. For the 1x dose condition, there was an interaction between step and talker bias ($\hat{\beta} = -1.430$, $SE = 0.623$, $z = -2.293$, $p = 0.022$), suggesting that the magnitude of the bias effect varied across continuum steps; this interaction was not reliable for the 2x dose condition ($\hat{\beta} = -0.962$, $SE = 0.544$, $z = -1.495$, $p = 0.135$).

To directly compare learning between the two dose conditions, an additional GLMM was constructed. The model structure was identical to that of the individual dose models except that dose (and all interactions with dose) were included in the fixed effects structure. Dose was entered into the model as a mean-centered contrast (1x = -0.5, 2x = 0.5). The full results of this model are shown in Table 3. Of note, the model showed a main effect of bias ($\hat{\beta} = 1.802$, $SE = 0.391$, $z = 4.613$, $p < 0.001$), but no significant interaction between bias and dose ($\hat{\beta} = -1.061$, $SE = 0.665$, $z = -1.595$, $p = 0.111$). Moreover, a likelihood ratio test showed no significant change in goodness of fit between the omnibus model and a simpler model in which dose was removed as a

and data visualization, and the interactions (Long, 2019) and cowplot (Wilke, 2019) packages that were used for data visualization.

fixed effect ($\chi(4) = 4.182$, $p = 0.382$).

Experiment 2

Consistent with Luthra et al. (2021), the results of experiment 1 confirm that listeners can simultaneously adapt multiple generative models given interleaved exposure to two talkers, at least for talkers of different genders. However – in contrast to Luthra et al. (2021) – we found no evidence that interleaved talker input required additional exposure beyond the standard dose in this paradigm (i.e., 20 critical exposures). Specifically, learning was observed in both the 1x and 2x dose conditions, and there was no evidence that the magnitude of learning differed between the two dose conditions. In experiment 1, though exposure to the two talkers was interleaved during the exposure phase, the test phase was blocked by talker. In experiment 2 we examine learning for interleaved talker input at *both* exposure and test. If listeners can dynamically retrieve distinct generative models, then learning will be observed even in the face of trial-by-trial talker variability at test. A failure to observe learning given interleaved test would suggest input-driven constraints on model retrieval.

Methods

Participants. Participants ($n = 80$) were recruited from the Prolific participant pool (Palan & Schitter, 2018) according to the criteria outlined for experiment 1. The sample included 42 women, 37 men, and one participant who preferred not to report gender. The mean age of the sample was 27 years ($SD = 5$ years). One additional participant was tested but excluded from analyses due to exhibiting a flat identification function at test. Participants were randomly assigned to either the 1x dose ($n = 40$) or 2x dose ($n = 40$) condition.

Stimuli. The stimuli were identical to those used in experiment 1.

Procedure. The procedure was identical to that described for experiment 1 with one key exception. Namely, instead of blocking test by talker, the 72 test trials (2 talkers x 6 continuum steps x 6 cycles) were presented in a single block that interleaved the two talkers' test stimuli. Specifically, six cycles of 12 test stimuli were presented, with each cycle consisting of different randomized order of both talkers' test stimuli.

Results

Performance was analyzed as outlined for experiment 1. Proportion correct talker identification during exposure was near ceiling (*mean* = 1.00, *SD* = 0.01, *range* = 0.94 – 1.00). Aggregate performance at test is shown in Figure 2. Visual inspection suggests a robust learning effect such that proportion *asi* responses are greater for the /s/-bias talker compared to the /ʃ/-bias talker. No effect of dose is visually apparent.

As shown in Table 2, the results of the models for each dose condition showed a robust influence of talker bias on *asi* responses. In both the 1x and 2x dose conditions, there were more *asi* responses for the /s/-bias talker compared to the /ʃ/-bias talker (1x: $\hat{\beta} = 2.246$, $SE = 0.529$, $z = 4.243$, $p < 0.001$; 2x: $\hat{\beta} = 1.944$, $SE = 0.483$, $z = 4.020$, $p < 0.001$). The interaction between step and bias was not significant for either dose condition ($p \geq 0.070$ in both cases).

The results of the model that included both dose conditions are shown in Table 3. There was a main effect of bias ($\hat{\beta} = 2.031$, $SE = 0.351$, $z = 5.793$, $p < 0.001$), consistent with the results of the individual models for each dose condition. There was no reliable interaction between talker bias and dose ($\hat{\beta} = 0.092$, $SE = 0.614$, $z = 0.150$, $p = 0.881$) nor between talker bias, dose, and continuum step ($\hat{\beta} = 0.339$, $SE = 0.588$, $z = 0.576$, $p = 0.564$), suggesting that the magnitude of the learning effect did not vary as a function of exposure dose. A likelihood ratio test showed no significant change in goodness of fit between the omnibus model and a simpler model in which dose was removed as a fixed effect ($\chi(4) = 1.422$, $p = 0.840$).

Experiment 3

The results of experiment 2 suggest that listeners can dynamically update, retrieve, and apply separate generative models for speech categorization even in the face of trial-by-trial talker variability in speech input. In experiment 3, we provide an even stricter test of this hypothesis by holding the critical fricative acoustics constant across talkers.

For interpreting the results of experiments 1 and 2, as is standard in this domain, we conclude that learning at test provides evidence that perception of acoustics (at test) has been conditioned on lexical context – and, here, talker identity – because of lexically-biased exposure. A limitation of this reasoning becomes apparent, however, when categorization includes speech from two different talkers; specifically, each step of the two test continua do not share identical acoustics. As we described previously, the manipulation used to create Peter's tokens resulted in a shift towards lower frequencies for the critical fricative acoustics in addition to the shift in fundamental frequency that was used to cue a male talker. Consequently, different categorization patterns may occur for the two talkers simply because of different fricative acoustics. In fact, a maximally parsimonious framework *should* categorize the speech of different talkers differently based on these acoustic differences and, indeed, perception of fricative acoustics is highly dependent on surrounding fundamental frequency (Johnson, 1991; Munson, 2011). As in Luthra et al. (2021), we mitigated this possibility in experiments 1 and 2 by counterbalancing the assignment of talker to each bias condition during exposure. However, the strongest test of the claim that learning is conditioned on lexical context (and, perhaps, talker) would be to present identical fricative acoustics across talkers, which we do in experiment 3. Given the different gender manipulation in this experiment, which is cued by a difference in fundamental frequency between the two talkers, equating fricative acoustics across talkers provides input that requires fricative identity and lexically guided learning to be conditioned on, at minimum, gender identity. Accordingly, examining learning under these conditions provides a strict test of the hypothesis that listeners update and retrieve multiple generative models for speech perception.

Methods

Participants. Participants (n = 80) were recruited from the Prolific participant pool (Palan & Schitter, 2018) according to the criteria outlined for experiment 1. The sample included 47 women and 33 men. The mean age of the sample was 27 years (*SD* = 5 years). Two additional participants were tested but excluded from analyses according to preregistered exclusion criteria including failure to pass the headphone screen (n = 1) and exhibiting a flat response function at test (n = 1). Participants were randomly assigned to either the 1x dose (n = 40) or 2x dose (n = 40) condition.

Stimuli. Stimuli for Joanne were identical to those used in experiments 1 and 2. Stimuli for Peter had one crucial difference. Specifically, while the non-fricative portions of all items were submitted to the Change Gender function as described in experiment 1, the fricatives were left unchanged. As shown in Figure 1, this resulted in fricative acoustics that were identical between the two talkers while still preserving source characteristics consistent with different gender talkers. To our ears, Peter's stimuli in experiment 3 were extremely difficult to perceptually discriminate from Peter's stimuli in experiments 1 and 2.

Procedure. The procedure was identical to that described for experiment 2; stimuli from the two talkers were interleaved at both exposure and test.

Results

Performance was analyzed as outlined for experiment 1. Proportion correct talker identification during exposure was near ceiling (*mean* = 0.99, *SD* = 0.02, *range* = 0.84 – 1.00).

Aggregate performance at test is shown in Figure 2. Visual inspection suggests a relatively weak learning effect in both dose conditions (compared to the learning effects observed in experiments 1 and 2), which appears to be slightly stronger in the 2x compared to the 1x dose condition.

The full results of the models for each dose condition are shown in Table 2. For the 1x dose condition, the main effect of bias was not statistically significant ($\hat{\beta} = 1.022$, $SE = 0.540$, $z = 1.892$, $p = 0.058$). However, there was a significant interaction between bias and continuum step ($\hat{\beta} = -0.849$, $SE = 0.413$, $z = 2.056$, $p = 0.040$), suggesting that the learning effect may be present at some but not all continuum steps. For the 2x dose condition, modest evidence for the main effect of bias was observed ($\hat{\beta} = 1.243$, $SE = 0.626$, $z = 1.987$, $p = 0.047$).

As shown in Table 3, when data from both dose conditions were examined together, a main effect of bias was observed ($\hat{\beta} = 1.121$, $SE = 0.415$, $z = 2.697$, $p = 0.007$). Though the p-value suggests a more robust effect compared to the individual dose models, we note that the magnitude of the effect (as quantified by the beta estimate) remains modest, consistent with the results of the individual dose models. There was no reliable interaction between talker bias and dose ($\hat{\beta} = 0.408$, $SE = 0.799$, $z = 0.511$, $p = 0.610$) nor between talker bias, dose, and continuum step ($\hat{\beta} = -0.027$, $SE = 0.558$, $z = -0.048$, $p = 0.962$), providing no evidence that magnitude of the learning effect varied as a function of exposure dose. A likelihood ratio test showed no significant change in goodness of fit between the omnibus model and a simpler model in which dose was removed as a fixed effect ($\chi(4) = 1.124$, $p = 0.891$).

Experiment 4

Collectively, the results of experiments 1 – 3 demonstrate that listeners can simultaneously learn and retrieve generative models for multiple talkers, even in the face of trial-by-trial talker variability at test, and even when the to-be-learned acoustics are identical across the two talkers. In these experiments, as in (to our knowledge) every other examination of mixed talker learning in this domain, the two talkers differed in gender. Though the results of experiments 1 – 3 are necessary to conclude that listeners engage in talker-specific perceptual learning, they are not sufficient given that using talkers of different genders introduces a confound between talker and gender. Indeed, the belief-updating framework under consideration here and invoked in Luthra et al. (2021) posits a hierarchy of generative models that may be used to guide adaptation, including language-specific models, gender-specific models, and talker-specific models.

The goal of experiments 4 – 6, which parallel experiments 1 – 3 as shown in Table 1, is to provide a critical test of talker-specific perceptual learning by presenting listeners with talkers who have perceptually distinct voices yet share the same gender. If learning is talker-specific, then the results of experiments 4 – 6 should yield the same patterns observed for experiments 1 – 3. A failure to observe the same learning patterns for talkers with perceptually distinct voices of the same gender would suggest that learning is linked to gender-specific instead of talker-specific models.

Methods

Participants. Participants (*n* = 80) were recruited from the Prolific participant pool (Palan & Schitter, 2018) according to the criteria outlined for experiment 1. The sample included 42 women and 38 men. The mean age of the sample was 29 years (*SD* = 5 years). Four additional participants were tested but excluded from analyses according to preregistered exclusion criteria including failure to pass the headphone screen (*n* = 1) and exhibiting a flat response function at test (*n* = 3). Participants were randomly assigned to either the 1x dose (*n* = 40) or 2x dose (*n* = 40) condition.

Stimuli. Stimuli for Joanne were identical to those of experiment 1. To create stimuli for a second female talker, referred to as Sheila, Joanne's stimuli were digitally manipulated using the Praat Vocal Toolkit (Corretge, n.d.). Specifically, a formant shift ratio of 0.8 was applied to simulate a change in vocal tract size, median pitch was set to 180 Hz to indicate a different sound source, and pitch variation was reduced by 10% to simulate different prosody. As for Peter's stimuli, stimulus duration was not altered. Perceptually, this process yielded tokens that cued a female talker with a voice that was perceptually distinct from Joanne.

Figure 3 displays two acoustic measurements for Joanne and Sheila's stimuli including fundamental frequency of the voiced portion of each token and center of gravity of the fricative portion of each token. Measurements were conducted as outlined for experiment 1. As can be viewed in Figure 3, (1) fundamental frequency was higher for Joanne compared to Sheila, (2) for both talkers, center of gravity for the clear variants is lower for /ʃ/ compared to /s/, and (3) center of gravity for the ambiguous variants falls intermediate to the clear variants. Figure 3 also shows that the digital signal manipulation used to change the sound source introduced a scaling of center of gravity in line with the change in fundamental frequency. Specifically, center of gravity for Sheila's tokens is shifted towards slightly lower frequencies compared to Joanne's tokens, yielding stimuli in which spectral characteristics for the critical /s/ and /ʃ/ energy in the tokens is not equated across talkers.

Procedure. The procedure was identical to that described for experiment 1, with the exception that Sheila's tokens were used in place of Peter's tokens. Accordingly, listeners heard tokens from Joanne and Sheila interleaved during the exposure phase, with the test phase blocked by talker.

Results

Performance was analyzed as outlined for experiment 1. Proportion correct talker identification during exposure was near ceiling (*mean* = 1.00, *SD* = 0.01, *range* = 0.96 – 1.00). Ceiling performance for talker identification during exposure confirms that the signal manipulation used to create Sheila's stimuli was sufficient to cue a perceptually distinct talker from Joanne.

Aggregate performance at test is shown in Figure 4. Visual inspection suggests no robust learning effect for either dose condition. This pattern was confirmed by GLMMs conducted for each dose condition. As shown in Table 4, no significant effect of talker bias was observed in either dose condition (1x: $\hat{\beta} = 0.198$, $SE = 0.446$, $z = 0.443$, $p = 0.658$; 2x: $\hat{\beta} = 0.770$, $SE = 0.420$, $z = 1.832$, $p = 0.067$). When data from both dose conditions were analyzed together, as reported in Table 5, no effect of talker bias was observed ($\hat{\beta} = 0.502$, $SE = 0.305$, $z = 1.643$, $p = 0.100$), consistent with the results of the individual models for each dose condition. Moreover, there was no reliable interaction between talker bias and dose ($\hat{\beta} = 0.419$, $SE = 0.520$, $z = 0.806$, $p = 0.420$), suggesting that learning did not vary as a function of exposure dose. Compared to a model that only included continuum step as a fixed effect, likelihood ratio tests did show a significant improvement in goodness of fit when bias was added to the model ($\chi(2) = 7.223$, $p = 0.027$), but no further improvement was observed when dose was additionally added to the model ($\chi(4) = 3.744$, $p = 0.442$).

Experiment 5

The results of experiment 4 provide no strong evidence of learning; though bias did improve model fit compared to a model that only predicted *asi* responses by continuum step, the bias effect was not significant either in the omnibus model or in the models that analyzed performance for each dose separately. These results are consistent with the interpretation that listeners adapt by updating gender-specific, and not talker-specific, models. That is, listeners may have updated a single model, linked to gender, that encompassed exposure from both women heard during exposure. Because lexical information differentially biased perception of the ambiguous fricative for each talker, linking the exposure input to a single model would result in no learning because of hearing the ambiguity in both biasing contexts. Indeed, Tzeng et al. (2021) observed no learning in a single talker condition in which listeners heard the talker produce the ambiguity in both biasing contexts, demonstrating that learning is linked to the consistency of the biasing input.

As described in the introduction, blocking the test phase by talker introduces a memory-based confound between talkers because the time between exposure and test cannot be held constant across talkers. In experiment 5, we examine whether learning for same gender talkers will emerge when speech from the two talkers is interleaved during exposure *and* test. If perceptual learning reflects updating of gender-specific models, as suggested by the results of experiment 4, then no learning will be observed in experiment 5. In contrast, evidence of learning in experiment 5 would suggest that listeners can maintain talker-specific models under specific circumstances, thus pointing to constraints on learning for same gender talkers.

Methods

Participants. Participants (n = 80) were recruited from the Prolific participant pool (Palan & Schitter, 2018) according to the criteria outlined for experiment 1. The sample included 44 women and 36 men. The mean age of the sample was 27 years (*SD* = 5 years). Two additional participants were tested but excluded from analyses according to preregistered exclusion criteria including failure to pass the headphone screen (n = 1) and exhibiting a flat response function at test (n = 1). Participants were randomly assigned to either the 1x dose (n = 40) or 2x dose (n =

40) condition.

Stimuli. The stimuli were identical to those used in experiment 4.

Procedure. The procedure was identical to that used in experiment 2, with the exception that Sheila's tokens were used in place of Peter's tokens. Specifically, Joanne and Sheila's tokens were interleaved at both exposure and test.

Results

Performance was analyzed as outlined for experiment 1. Proportion correct talker identification during exposure was near ceiling (*mean* = 0.99, *SD* = 0.01, *range* = 0.92 – 1.00). Aggregate performance at test is shown in Figure 4. Visual inspection suggests a learning effect reflecting more *asi* responses for the /s/-bias talker compared to the /ʃ/-bias talker. No effect of dose is visually apparent.

GLMMs for each dose condition (Table 4) confirmed a significant effect of talker bias in each of the 1x dose ($\hat{\beta}$ = 1.442, *SE* = 0.428, *z* = 3.369, *p* = 0.001) and 2x dose ($\hat{\beta}$ = 1.041, *SE* = 0.434, *z* = 2.399, *p* = 0.016) conditions. As shown in Table 5, a main effect of talker bias was also observed when data from the two dose conditions were analyzed together in a single model; however, no significant interaction between talker bias and dose ($\hat{\beta}$ = 0.148, *SE* = 0.533, *z* = 0.277, *p* = 0.782) or between talker bias, dose, and continuum step ($\hat{\beta}$ = -0.391, *SE* = 0.390, *z* = -1.003, *p* = 0.316) was observed. Furthermore, a likelihood ratio test showed no significant change in goodness of fit between the omnibus model and a simpler model in which dose was removed as a fixed effect (χ(4) = 1.654, *p* = 0.799).

Experiment 6

The results of experiment 5 suggest that listeners can update and retrieve talker-specific generative models given that learning was observed for two talkers who were members of the same sociophonetic class (here, gender). However, in context of the null learning effect observed in experiment 4, the results of experiment 5 suggest that maintaining distinct models for same gender talkers may be more fragile than maintaining distinct models for different gender talkers, a point we elaborate on further in the discussion. Experiment 6 provides an additional test of talker-specificity for same gender talkers, following the manipulation of experiment 3 in which the critical fricative acoustics are held constant between the two talkers.

Methods

Participants. Participants (n = 80) were recruited from the Prolific participant pool (Palan & Schitter, 2018) according to the criteria outlined for experiment 1. The sample included 47 women, 32 men, and one participant who declined to report gender. The mean age of the sample was 26 years (SD = 5 years). Three additional participants were tested but excluded from analyses according to preregistered exclusion criteria including failure to pass the headphone screen (n = 1) and exhibiting a flat response function at test (n = 2). Participants were randomly assigned to either the 1x dose (n = 40) or 2x dose (n = 40) condition.

Stimuli. The stimuli for Joanne were identical to those used in all previous experiments. Stimuli for Sheila had one crucial difference. Specifically, while the non-fricative portions of all items were modified as described for experiment 4, the fricatives were left unchanged. As shown in Figure 3, this resulted in fricative acoustics that were identical between the two talkers while still preserving source characteristics consistent with two female talkers. To our ears, Sheila's stimuli for experiment 6 were extremely difficult to perceptually discriminate from Sheila's stimuli for experiments 4 and 5, though Sheila's stimuli were readily discriminable from Joanne's stimuli in all cases.

Procedure. The procedure was identical to that described for experiment 5; stimuli for the two talkers were interleaved at both exposure and test.

Results

Performance was analyzed as outlined for experiment 1. Proportion correct talker identification during exposure was near ceiling (*mean* = 0.99, *SD* = 0.01; *range* = 0.94 = 1.00). Aggregate performance at test is shown in Figure 4. Visual inspection suggests a minimal learning effect that appears to be slightly stronger in the 2x compared to the 1x dose condition. The results of the GLMM for each dose condition are shown in Table 4. The effect of talker bias was not significant in the 1x dose condition ($\hat{\beta}$ = 0.319, *SE* = 0.405, *z* = 0.787, *p* = 0.431) but it did meet threshold for statistical significance in the 2x dose condition ($\hat{\beta}$ = 0.904, *SE* = 0.451, *z* = 2.005, *p* = 0.045). The results of the GLMM aggregating across dose conditions is shown in Table 5. A modest effect of talker bias was observed ($\hat{\beta}$ = 0.592, *SE* = 0.300, *z* = 1.973, *p* = 0.048), reflecting more *asi* responses for the /s/-bias talker compared to the /ʃ/-bias talker. Neither the interaction between talker bias and dose ($\hat{\beta}$ = 0.267, *SE* = 0.580, *z* = 0.460, *p* = 0.646) nor the interaction between talker bias, dose, and continuum step ($\hat{\beta}$ = 0.246, *SE* = 0.467, *z* = 527, *p* = 0.598) was reliable, and a likelihood ratio test showed no significant change in goodness of fit between the omnibus model and a simpler model in which dose was removed as a fixed effect (χ(4) = 3.125, *p* = 0.537).[2]

[2] In keeping with best practices for promoting reproducibility of research, we have reported all experiments conducted for this project in the main text except one, which we report in this footnote. Selective reporting of experiments in the scientific literature (e.g., running numerous experiments around a central hypothesis and selectively reporting only the ones that "worked") is a questionable, and, unfortunately, common research practice (John et al., 2012) that contributes to reduced reproducibility of research by eliminating important context needed to interpret reported findings in the scientific literature and because versions that "worked" may reflect false positives in the context of multiple related studies that did not yield reliable effects (Greenwald, 1975; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). In addition to the six experiments reported in the main text, we ran one pilot experiment that was identical to experiment 1 except that the test phase included two additional tokens for each talker's

Effect sizes across experiments

Learning in the lexically guided perceptual learning paradigm is often considered as a binary outcome (e.g., learning occurs or it does not occur, Luthra et al., 2021); however, a growing body of evidence suggests that lexically guided perceptual learning may be more optimally characterized as a graded outcome that is linked to characteristics of the specific acoustic input presented during exposure (Drouin et al., 2016; Tzeng et al., 2021). Though linking learning to the specific spectral patterns presented during exposure is beyond the scope of the current work, Figure 5 shows the learning effect sizes for 11 lexically guided perceptual learning studies, including the six experiments presented here, the four interleaved exposure experiments in Luthra et al. (2021), and experiment 1A from Tzeng et al. (2021). We provide this effect size comparison to promote consideration of learning as a graded outcome that may vary across stimulus sets, which is an important avenue for future research. Recall that Joanne's stimuli were identical to those used in Tzeng et al. (2021), which also used the standard lexically guided perceptual learning paradigm (i.e., listeners were only exposed to one talker and learning was assessed as a between-subjects effect, with different listeners exposed to either Joanne's /s/-bias or /ʃ/-bias stimuli).

Visual inspection of Figure 5 shows wide variability in effect sizes between the current experiments and those of Luthra et al. (2021), especially for the 1x dose condition. As we discuss further in the discussion section, the weak learning magnitude in the 1x dose conditions of Luthra et al. (2021) may explain, at least in part, why the current experiments failed to replicate

continuum, one near each of the continuum endpoints. The same pattern of results was observed in the pilot experiment as was observed for experiment 1. Based on the results of the pilot experiment, we decided to narrow the range of the test continuum to better sample perception in the more intermediate region where learning was predicted to occur, as described for the six experiments presented in the main text.

the finding that simultaneously adapting to multiple talkers requires additional exposure beyond the standard exposure dose. Numerically, the magnitude of the learning effect is in most cases larger for the different gender experiments (experiments 1 – 3) compared to the same gender experiments (experiments 4 – 6) in the current work. Of note, the magnitude of the learning effect for some of the mixed talker experiments presented here falls comfortably in the confidence interval of the effect size observed in Tzeng et al., suggesting that under some circumstances, simultaneously adapting to two talkers can lead to learning of a similar magnitude as adapting to a single talker. Visual inspection of Figure 5 also suggests that the magnitude of the learning effect was not constant even among conditions that held talker gender constant during exposure. For example, the magnitude of the learning effect for the 1x dose conditions is larger in experiments 1 and 2 compared to experiment 3; though all three of these experiments presented talkers of different genders during exposure. In the following sections, we describe a series of exploratory analyses that formally examined the magnitude of the learning effect across the six experiments reported in this manuscript.

Exploratory analysis: Comparing learning within gender conditions

A set of analyses was conducted to compare the magnitude of the learning effect across experiments 1 – 3 (the different gender experiments) and, separately, across experiments 4 – 6 (the same gender experiments). We explicitly note that these analyses should be considered exploratory given that the sample size was not planned to detect potential interactions between learning and experiment. We present the full exposition of these analyses in the Supplementary Material and summarize the key findings here in the main text.

Different gender experiments

We first examined learning across experiments separately for each dose. For the 1x dose condition, a significant effect of bias was observed ($\hat{\beta}$ = 1.854, *SE* = 0.310, *z* = 5.979, *p* < 0.001), but there was no strong evidence to suggest that the magnitude of the bias effect differed between experiments 1 and 2 ($\hat{\beta}$ = -0.375, *SE* = 0.698, *z* = -0.538, *p* = 0.591) or between experiments 2 and 3 ($\hat{\beta}$ = -0.998, *SE* = 0.663, *z* = -1.506, *p* = 0.132). The same pattern held for the 2x dose condition, which showed a reliable effect of bias ($\hat{\beta}$ = 1.491, *SE* = 0.326, *z* = 4.569, *p* < 0.001) but no significant interaction between bias and experiment for either the experiment 1 versus experiment 2 contrast ($\hat{\beta}$ = 0.696, *SE* = 0.758, *z* = 0.917, *p* = 0.359) or the experiment 2 versus experiment 3 contrast ($\hat{\beta}$ = -0.550, *SE* = 0.739, *z* = -0.743, *p* = 0.457). The same pattern held when data from both dose conditions were analyzed together (p ≥ 0.117 in both cases).

Same gender experiments

The same procedure was used to compare the magnitude of the learning effect across experiments 4 – 6. For both the 1x and 2x dose conditions, there was a significant effect of bias (1x: $\hat{\beta}$ = 0.642, *SE* = 0.243, *z* = 2.648, *p* = 0.008; 2x: $\hat{\beta}$ = 0.899, *SE* = 0.265, *z* = 3.395, *p* = 0.001), but no robust evidence to suggest that learning differed between experiment 4 and experiment 5 or between experiment 5 and experiment 6 (*p* > 0.106 in all cases). When data from both dose conditions were analyzed together, the experiment by bias interaction coefficients provide marginal evidence to suggest that learning increased from experiment 4 to experiment 5 ($\hat{\beta}$ = 0.684, *SE* = 0.413, *z* = 1.656, *p* = 0.098) and decreased from experiment 5 to experiment 6 ($\hat{\beta}$ = -0.690, *SE* = 0.395, *z* = -1.745, *p* = 0.081).

Exploratory analysis: Comparing learning across gender conditions

The results of the individual experiments suggest that under some circumstances, perceptual learning for speech reflects dynamic updating and retrieval of talker-specific generative models. Specifically, the results of experiment 5 showed a robust learning effect given exposure to two talkers of the same sociophonetic class (i.e., gender, here). However, learning was not observed for the same gender talkers when test was blocked by talker (experiment 4). Moreover, visual inspection of Figure 5 suggests that learning effect sizes in the same gender experiments (i.e., experiments 4 – 6) are, in general, weaker than the effect sizes observed in the different gender experiments (i.e., experiments 1 – 3). To examine whether the magnitude of learning differed as a function of exposure to different gender versus same gender talkers, exploratory analyses were conducted to compare learning between parallel experiments (i.e., experiment 1 vs. experiment 4, experiment 2 vs. experiment 5, experiment 3 vs. experiment 6). The full exposition of these analyses is provided in the Supplementary Material; we summarize the key findings here.

For the 1x dose conditions, there was a significant interaction between gender match and bias when test was blocked by talkers ($\hat{\beta} = -1.775$, $SE = 0.533$, $z = -3.330$, $p = 0.001$), indicating a larger learning effect when the two exposure talkers differed in gender (experiment 1) compared to when they were the same gender (experiment 4). The interaction between gender match and bias was not reliable when comparing experiment 2 to experiment 5, nor when comparing experiment 3 to experiment 6 ($p > 0.106$ in both cases). For the 2x dose conditions, no significant interactions between gender match and bias were observed ($p > 0.388$ in all cases). When data from both dose conditions were analyzed together, there was a significant interaction between gender match and bias only when test was blocked by talker (i.e., learning was larger in experiment 1 compared to experiment 4; $\hat{\beta} = -1.047$, $SE = 0.415$, $z = -2.524$, $p = 0.012$), consistent with the results of the individual dose models.

Exploratory analysis: Reaction time at test

As described in the introduction, previous research suggests a processing cost associated with simultaneous adaptation to multiple talkers (Luthra et al., 2021). The evidence used to support this conclusion reflected a null learning effect (i.e., no significant effect of lexical bias) when interleaved exposure to two talkers' speech consisted of the standard exposure dose, and a significant learning effect when dose was doubled. Though a significant interaction between learning and dose was not observed in this study, this pattern of results is broadly consistent with the conclusion that twice as much exposure is required for learning to occur given interleaved compared to blocked mixed talker exposure (Luthra et al., 2021). However, the exact locus of this processing cost is unknown, and it was not replicated in the current experiments.

Based on research suggesting that mixed talker input results in increased processing time compared to single talker input due to disruptions in auditory attention (e.g., Choi & Perrachione, 2019) and the need for talker normalization (Saltzman et al., 2021), we performed an exploratory analysis on reaction time at *test*, capitalizing on the manipulations of the current study that included both blocked and interleaved test. Given that we did not find any evidence interleaved *exposure* impeded the emergence of the learning effect, nor any evidence that exposure dose impacted learning, the goal of this analysis was to help elucidate the nature of a processing cost, if any, associated with interleaved *test*.

To visualize RT, mean RT was first calculated for each participant, collapsing across all test trials. Grand means for each experiment were then calculated over by-subject means separately for the two dose conditions. Figure 6 shows mean RT in each experiment separately for each dose. Visual inspection suggests two patterns. First, RTs between the dose conditions for a given experiment show minimal difference. Second, RTs show a monotonic increase in line with increased talker variability *and* increased conditioning of phonetic variation on talker. That is, for each gender match condition, mean RT is slower in when test was interleaved across talkers (experiment 2, experiment 5) compared to when test was blocked by talker (experiment 1, experiment 4). This increase in RT is consistent with previous research demonstrating slower processing time in mixed talker compared to single talker input (e.g., Choi & Perrachione, 2019; Magnuson & Nusbaum, 2007; Saltzman et al., 2021; Stilp & Theodore, 2020). When test was interleaved *and* the fricative acoustics were held constant across talkers, RTs slow even further. Because experiments 2 and 3 (and the parallel same gender examinations, experiments 5 and 6) both used interleaved test, increased RTs here may reflect increased computational complexity in resolving phonetic identity given identical fricative acoustics across talkers.

To examine these patterns statistically, trial-level raw RTs were submitted to a generalized linear mixed effects model with the Gamma response family and an identity link function following recommendations of Lo and Andrews (Lo & Andrews, 2015). The model structures were parallel and included fixed effects of gender match, dose, experiment, and their interactions. Gender match (different gender = -0.5, same gender = 0.5) and dose (1x = -0.5, 2x = 0.5) were sum-coded. Experiment was entered into the model as two sliding contrasts, one that compared blocked test (experiments 1 and 4) to interleaved test (experiments 2 and 5; E1/E4 = -2/3, E2/E5 = 1/3, E3/E6 = 1/3), and one that compared interleaved test when fricative acoustics differed between talkers (experiments 2 and 5) to interleaved test when fricatives acoustics were held constant between talkers (experiments 3 and 6; E1/E4 = -1/3, E2/E5 = -1/3, E3/E6 = 2/3). The random effects structure consisted of random intercepts by subject and random intercepts by continuum step. The results of this model are shown in Table 6. Reaction time significantly increased from blocked test to interleaved test ($\hat{\beta} = 48.160$, $SE = 17.623$, $z = 2.733$, $p = 0.006$) and between the two interleaved test manipulations ($\hat{\beta} = 56.881$, $SE = 17.741$, $z = 3.206$, $p = 0.001$). No other main effects or interactions were reliable, including the main effect of gender match condition ($p \geq 0.114$ in all cases).

Discussion

The current investigation examined listeners’ ability to simultaneously learn multiple talkers’ idiolects. In each of six experiments, listeners were exposed to the speech of two talkers, with productions from the two talkers randomly interleaved during exposure. Lexical context was used to bias listeners to perceive an ambiguous fricative as /s/ for one talker and /ʃ/ for the other talker. During exposure, listeners completed a talker identification task. Talker identification accuracy for all experiments was near ceiling, confirming that listeners perceived the speech as being produced by two different talkers. Following exposure, listeners categorized tokens drawn from two *ashi* – *asi* continua, one for each talker. The manipulations within and across the six experiments were designed to inform four facets of perceptual learning for multiple talkers, including (1) flexibility in updating and applying talker-specific models, (2) gender- versus talker-specificity, (3) necessity of conditioning perception of acoustics on a talker, and (4) perceptual costs associated with simultaneous maintenance of multiple models.

While experiment 1 blocked the categorization test by talker, experiment 2 interleaved test tokens randomly across talkers, potentially requiring more frequent shifts in the models used to inform categorization. Experiment 3 additionally manipulated the acoustics of input such that the fricatives of the two talkers were physically identical across talkers. The second set of experiments (experiments 4 – 6) mirrored experiments 1 – 3 except to present listeners with talkers of the same gender during exposure. Finally, all six experiments examined learning as a function of two exposure doses to provide a replication of the reported perceptual cost of simultaneously updating multiple talker-specific models (Luthra et al., 2021). The implications of the results for each of these manipulations towards a theory of perceptual learning for multiple talkers are discussed in turn, below.

As reviewed in the introduction, the extant literature provides mixed evidence as to the specificity of perceptual learning given exposure to a single talker's idiosyncratic productions, with generalization to a novel talker observed in some cases but not in others (Eisner & McQueen, 2005; Tamminga, et al., 2020; Kraljic & Samuel, 2005). More consistent support for talker-specific learning comes from investigations of mixed talker listening environments (Kraljic & Samuel, 2007; Luthra, et al., 2021), including the results of the current work. Because past research consistently blocked test by talker, even when exposure was provided to multiple talkers, it did not examine listeners' ability to dynamically retrieve talker-specific models. Indeed, an inherent consequence of context-specific versus a more parsimonious representational structure is the necessity to change which model is being used each time the source input changes. The results of the current experiments demonstrate that trial-by-trial talker variability is not a constraint on model retrieval. Learning was observed in experiments 2 and 5 (and, perhaps, to a lesser extent, experiments 3 and 6), which interleaved talkers' speech at test. These findings suggest a perceptual system that can dynamically shift between different generative models on a trial-by-trial basis. This is crucially orthogonal to the question of talker-specific adaptation in single-talker environments; a mechanism without dynamic switching capability could be forced to generalize across talkers in mixed talker environments regardless of its specificity in single talker contexts.

Between-talker similarity has been implicated both theoretically and empirically as a determinant of generalization across talkers (Kraljic & Samuel, 2005; Kleinschmidt & Jaeger, 2015). Specifically, if within-talker variability of individual talkers outweighs between-talker differences, then there is no predicted gain to maintaining talker-specific models. As a reductio ad absurdum, consider two talkers who are indexically distinct but whose phonetic patterns are identical – even the most specific of perceptual mechanisms should generalize between these talkers. While encountering such a pair of talkers is highly unlikely, a more moderate corollary is indeed quite common – talkers who share a salient indexical trait, in our case gender, and who therefore are acoustically less distinct than talkers of different genders. In the current work, talker-specific learning was observed for both different gender and same gender talkers. However, learning for different gender talkers was more robust than for same gender talkers given that it was maintained in the face of blocked test in the former but not the latter. This is perhaps explained by an interleaved test phase highlighting between-talker differences, as these differences are encountered on a trial-by-trial basis given interleaved test. Blocked test (experiments 1 and 4), in contrast, may make between-talker differences less salient, leading the system to generalize rather than maintain talker-specific learning. Though the magnitude of the learning effect (as indexed by the beta estimates for the fixed effect of bias in the regression models, shown in Figure 5) were numerically lower for the same gender experiments (experiments 4 – 6) compared to their different gender counterparts (experiments 1 – 3), the only case in which robust evidence of attenuated learning for same compared to different gender talkers was observed was when test was blocked by talker.

The current results are consistent with a theory that predicts interleaved experience as *more* conducive to learning. While a priori hypotheses focused on dynamically shifting between models as a potentially costly process, a less obvious benefit of frequent shifting is that each model is retrieved more often. Regular reactivation may offset the possibility of memory decay; one is unlikely to forget a model in frequent use. If memory decay is indeed a risk in mixed talker learning conditions, then frequent activation of talker-specific models can be viewed as conducive to learning. On the one hand, previous research has shown that given exposure to a single talker, learning is robust to both time (i.e., it persists 12 hours after exposure) and exposure to speech from other talkers (Eisner & McQueen, 2006; Kraljic & Samuel, 2005). On the other hand, previous research has shown that given exposure to a single talker, learning is attenuated even in the course of a single test phase, presumably reflecting distributional learning that occurs given exposure to the test stimuli themselves (Giovannone & Theodore, 2021; Liu & Jaeger, 2018, 2019; Tzeng et al., 2021). Future research is needed not only to reconcile these disparate results for single talker learning, but also to integrate findings on single talker learning environments with those of mixed talker learning environments. Moreover, future research is needed to determine if the current results generalize to mixed talker environments with more than two talkers.

In the current work, as in the broader lexically guided perceptual learning domain, evidence for learning is measured by a difference in categorization at test following lexically biased exposure. When different continua are used to assess learning, as must occur to some degree when learning is assessed for multiple talkers, it is important to consider an alternative explanation for differences in categorization between test continua, which is that they may reflect differences in the acoustics of the test continua, orthogonal to specificity of learning. We sought to mitigate this possibility in several ways. First, following Luthra et al. (2021), talkers were counterbalanced across biasing conditions in each experiment. Additionally, experiment 3 and experiment 6 eliminated acoustic differences between the fricatives of the two talkers by using the same physical fricatives for both talkers. Crucially, this was not designed to facilitate identical response functions between the talkers, given that between-talker differences in fundamental frequency both heavily condition perception of fricative identity (Johnson, 1991;

Munson, 2011) and were necessary in our experiments to perceptually cue two different talkers during exposure. Rather, by using identical fricatives, we provided a more robust test of the guiding framework, which is that differences in categorization manifest from differences in conditional interpretation of the input based on talker-specific patterns. Based on the results of the individual experiments, learning does appear to be challenged when fricative acoustics were identical across talkers. As shown in Tables 3 and 5 (and Figure 5), the magnitude of the bias effect in experiment 2 (β = 2.031, $p < 0.001$) is approximately twice the magnitude of the bias effect in experiment 3 (β = 1.121, p = 0.007). Likewise, the magnitude of the bias effect in experiment 5 is (β = 1.221, $p < 0.001$) is approximately twice the magnitude of the bias effect in experiment 6 (β = 0.592, p = 0.048). These results suggest that learning may have been diminished when fricatives were identical between talkers (experiments 3 and 6) compared to when they were allowed to exhibit more natural variation (experiments 2 and 5), though we note that the exploratory analyses that compared learning across experiments (reported in full in the Supplementary Material) provided no strong evidence for these interactions.

Luthra et al. (2021) describe a perceptual cost to learning the distributions of two talkers simultaneously; namely, that twice as much exposure is required for mixed talker compared to single talker learning. Seeking to replicate this cost, the six experiments reported here examined learning for both a standard exposure dose (20 ambiguous exposures in lexically biasing contexts for each talker) and for twice the standard exposure dose (40 ambiguous exposures in lexically biasing contexts for each talker). None of the six experiments found strong evidence to suggest that learning was influenced by exposure dose. Here we consider three potential explanations for why we did not replicate this processing cost. First, the exact dose in Luthra et al. (2021) was 16 or 32 critical exposures for each talker in the 1x and 2x dose conditions, respectively. It may be the case that the additional four exposures provided in the current 1x dose conditions were sufficient to overcome the processing cost observed in Luthra et al. (2021). Presumably – even for single talker environments – there is a relationship between exposure dose and learning. Indeed, Tzeng et al. (2021) found that the magnitude of learning for a single talker was larger given 20 compared to 10 critical exposures. The ideal adapter framework invoked here hypothesizes an influence of exposure dose for adaptation; namely, this framework predicts that evidence from a given talker must be sufficiently deviant to trigger the updating of an existing generative model and that initial evidence of a deviation is weighted more strongly than subsequent evidence that conforms to the deviation (Kleinschmidt & Jaeger, 2015). On this view, a difference between 16 and 20 critical exposures may in principle lead to different model beliefs, which would in turn impact observed learning. A fruitful avenue for future research would be to test these predictions more explicitly by parametrically manipulating exposure dose and number of exposure talkers at a fine grain.

Second, the disparate results between Luthra et al. (2021) and the current work may reflect differences in the specific stimuli used to elicit and measure learning. Informal reports suggest that learning in the lexically guided perceptual learning paradigm shows wide variability across stimulus sets. As shown in Figure 5, the magnitude of learning across stimulus sets used in the current work and Luthra et al. (2021) shows wide variability. Specifically, the effect size of the learning effect in the single dose conditions of Luthra et al. (2021) are relatively small compared to those of the current investigation. This could be indicative of stimuli that are more difficult for listeners to learn from and thus generate talker-specific models for, and therefore yield a potential specific situation in which additional exposure is beneficial. That is, if aspects of the stimuli are not conducive to strong learning in general, then learning may be more sensitive to task-based changes. Future research is needed to better explicate the relationship between exposure (and test) acoustics and subsequent learning, including examining whether the patterns observed here generalize to other phonological contrasts, which could potentially be used to explain the different learning magnitude elicited across different stimulus sets in the lexically guided perceptual learning domain. That there is no apparent cost of interleaving input from the two talkers in exposure in the current work converges with the lack of a cost found in our results for interleaved versus blocked talker input at test. Collectively, the current findings indicate that listeners can dynamically shift between talker-specific models without requiring additional exposure to simultaneously learn two talkers' idiosyncratic speech patterns and without an attenuation of learning given trial-by-trial talker variability at test. The current investigation is of course limited in that we did not include any manipulations where the *exposure* phase was blocked by talker, and therefore cannot make strong claims as to the potential role of interleaved exposure for our specific stimuli.

Third, we may not have observed robust statistical evidence for an influence of exposure dose on learning because our experiments may have been underpowered to detect an interaction of this magnitude. Designing the current experiments to have power to detect this interaction was challenged because the bias by dose interaction was not statistically reliable in Luthra et al. (2021). A priori power analyses are generally conducted to detect effect sizes that meet threshold for statistically significance; that is, null effects generally do not meet criteria for an effect size *of interest*. Accordingly, the sample size for the current experiments was set to detect an effect of bias of the magnitude observed in Luthra et al. in the experiments where learning was reported (that also yielded power to detect a bias by dose interaction of the magnitude observed in Tzeng et al. for single talker exposure), and we analyzed performance for each dose condition separately (in addition to testing for the bias by dose interactions). However, we did conduct a power analysis based on the magnitude of the (null) bias by dose interaction observed in Luthra et al. (2021) using the same simulation procedure described for the power analysis reported in experiment 1. The results showed that 512 participants, 256 in each dose condition, would be required to have adequate power (82%) to detect an interaction of that magnitude. Though the sample size in the current experiments (*n* = 40 in each between-subject condition) is slightly higher than the sample size in Luthra et al. (*n* = 32), both sample sizes are far below the sample size indicated by the power analysis. An informal analysis of more than 100 previous studies in the standard lexically guided perceptual learning literature revealed wide variability in sample sizes, including as a few as 12 subjects (e.g., Drouin et al., 2016; Mitterer & Reinisch, 2013; Myers & Mesite, 2014) to a maximum of 63 subjects in each between-subjects condition (Nelson & Durvasula, 2021), with an approximate mean sample size of 26 subjects (*SD* = 12 subjects; *median* = 24 subjects) in each between-subjects condition. One way to interpret the sample size convention in this domain is that it reflects what the field has collectively determined to be an effect size "of interest," consistent with the convention to assess learning outcomes under specific, individual experimental situations. As the field moves to consider learning as a graded instead of a binary outcome (e.g., Tzeng et al., 2021), future investigations may require more explicit specification of effect sizes of interest and a revision to the sample size convention in this domain to ensure that experiments are adequately powered to detect explicitly identified effect sizes of interest. Though conducting informative, appropriate power analyses has been a longstanding challenge for research in general, widespread availability of open access data combined with new tools for conducting power analyses (e.g., Green & MacLeod, 2016) mitigates some of the traditional challenges in conducting a priori power analyses. We note that code to reproduce the power analyses presented in this manuscript is available on the OSF repository for this manuscript (https://osf.io/x4yeq/); this code is heavily commented to promote its reuse.

Though we find no evidence of a processing cost for mixed talker exposure in terms of dose required to support learning or the magnitude of the learning effect, we did observe systematic increases in the processing time for phonetic categorization at test; processing time increased as a function of both increased talker variability (i.e., single vs. mixed talkers) and increased conditioning of phonetic variation on talker, as shown in Figure 6. This pattern of processing time is in line with accounts of the role of auditory attention (Choi & Perrachione, 2019; Kapadia & Perrachione, 2020) and talker normalization (Magnuson et al., 2021; Magnuson & Nusbaum, 2007; Saltzman et al., 2021) on phonetic identification. These accounts may interact with each other to the extent that if frequent changes in talker disrupt auditory attention, then this could in turn reduce resources available for resource-demanding, talker-specific computations that bring talker identity and phonetic information into alignment for speech sound categorization. Specifically, increased processing time when the talkers' speech was interleaved at test compared to when test was blocked by talker is predicted by both auditory attention and talker normalization frameworks because of trial-by-trial talker variability in the former but not the latter. The further increase in processing time when acoustics were held constant across talkers could be accommodated by the contextual tuning theory of talker normalization. Phonetic constancy across talkers predicts slowed processing if the time to compute a talker-specific mapping is graded to reflect increased challenge in aligning talker and phonetic information over identical compared to distinct phonetic cues.

Finally, the current results provide insight towards further specification of the ideal adapter framework (Kleinschmidt & Jaeger, 2015), which has increasingly been invoked as an explanatory theory for speech adaptation (Clayards et al., 2008; Giovannone & Theodore, 2021; Kleinschmidt, 2019; Kleinschmidt et al., 2015; Kleinschmidt & Jaeger, 2016; Luthra et al., 2021; Saltzman & Myers, 2021; Theodore et al., 2019; Theodore & Monto, 2019; Tzeng et al., 2021; Xie et al., 2018, 2021). Consistent with this framework, the current results found evidence that listeners can, in some circumstances, maintain distinct generative models for individual speakers. In its current instantiation, the ideal observer framework remains agnostic regarding potential processing costs incurred by switching the between generative models on successive utterances. The current findings support a framework where such a cost is not evident. If anything, the current results show that simultaneous learning of multiple talkers, at least when they are of the same gender, is facilitated by their speech being interleaved in time, necessitating frequent model switching. The ideal adapter framework additionally posits that learning should be conditioned heavily on the "situation" in which input occurs, which predicts equally robust learning when fricative acoustics are held identical between talkers compared to when they vary across talkers, provided that the talkers themselves (i.e., the "situation" in this framework) can be discriminated. That is, identical acoustics in different situations should not necessarily be more difficult to learn than different acoustics in different situations according to the ideal adapter framework. However, our findings suggest that learning for multiple talkers may be attenuated when fricative acoustics are held constant between talkers, suggesting that the ideal adapter framework may need refinement to account for the current results.

Conclusions

The present study informs theories of perceptual learning in mixed talker listening environments. Consistent with previous research, we observed evidence of a learning mechanism that is capable of simultaneously updating multiple generative models. Extending past research, we found evidence that (1) listeners can dynamically apply talker-specific models for speech categorization when faced with trial-by-trial talker variability in speech input, (2) frequent changes to model retrieval do not attenuate learning, (3) simultaneously updating talker-specific models for multiple talkers does not require increased exposure, and (4) between-talker similarity, including a shared identical trait, may under some circumstances constrain talker-specific learning. The current results collectively suggest that perceptual learning for speech is achieved via a mechanism that represents a context-dependent, cumulative integration of experience with speech input and can dynamically apply different generative models in mixed talker listening environments.

Acknowledgements

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

Open Practices Statement

Preregistration, stimuli, trial-level data, analysis code, and Supplementary Material are available at https://osf.io/x4yeq/.

References

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

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

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

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

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

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

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

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.

Choi, J. Y., & Perrachione, T. K. (2019). Time and information in perceptual adaptation to speech. *Cognition*, *192*, 103982.

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

Clopper, C. G., & Pisoni, D. B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech*, *47*(3), 207–238.

Corretge, R. (n.d.). *Praat Vocal Toolkit*. http://www.praatvocaltoolkit.com

DiCanio, C. (n.d.). Retrieved May 15, 2022, from https://www.acsu.buffalo.edu/~cdicanio/scripts/Time_averaging_for_fricatives_4.0.praat

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

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

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

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*(4), 1950–1953.

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

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

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

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20.

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

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

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Johnson, K. (1991). Differential effects of speaker and vowel variability on fricative perception. *Language and Speech*, *34*(3), 265–279.

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

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

Kapadia, A. M., & Perrachione, T. K. (2020). Selecting among competing models of talker adaptation: Attention, cognition, and memory in speech processing efficiency. *Cognition*, *204*, 104393.

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

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

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*, *34*(1), 43–68.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

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

Kleinschmidt, D. F., Raizada, R. D., & Jaeger, T. F. (2015). Supervised and unsupervised learning in phonetic adaptation. *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*.

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

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

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

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

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

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

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

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171.

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

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

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391–409.

Magnuson, J. S., Nusbaum, H. C., Akahane-Yamada, R., & Saltzman, D. (2021). Talker familiarity and the accommodation of talker variability. *Attention, Perception, & Psychophysics*, *83*(4), 1842–1860.

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

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

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

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

Mitterer, H., & Reinisch, E. (2013). No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. *Journal of Memory and Language*, *69*(4), 527–545.

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

Myers, E. B., & Mesite, L. M. (2014). Neural systems underlying perceptual adjustment to non-standard speech tokens. *Journal of Memory and Language*, *76*, 80–93.

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

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

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science | Science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Owren, M. J. (2008). GSU Praat Tools: Scripts for modifying and analyzing sounds using Praat acoustics software. *Behavior Research Methods*, *40*(3), 822–829. https://doi.org/10.3758/BRM.40.3.822

Palan, S., & Schitter, C. (2018). Prolific. Ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530.

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

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

Saltzman, D., Luthra, S., Myers, E. B., & Magnuson, J. S. (2021). Attention, task demands, and multitalker processing costs in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(12), 1673–1680.

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

Samuel, A. G. (2016). Lexical representations are malleable for about one second: Evidence for the non-automaticity of perceptual recalibration. *Cognitive Psychology*, *88*, 88–114. https://doi.org/10.1016/j.cogpsych.2016.06.007

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

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

Stilp, C. E., & Theodore, R. M. (2020). Talker normalization is mediated by structured indexical information. *Attention, Perception & Psychophysics*, *82*(5), 2237–2243.

Tarabeih-Ghanayim, M., Lavner, Y., & Banai, K. (2020). Tasks, talkers, and the perceptual learning of time-compressed speech. *Auditory Perception & Cognition*, *3*(1–2), 33–54.

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

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

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

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

Theodore, R. M., Myers, E. B., & Lomibao, J. A. (2015). Talker-specific influences on phonetic category structure. *The Journal of the Acoustical Society of America*, *138*(2), 1068–1078.

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

van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1483–1494. https://doi.org/10.1037/0096-1523.33.6.1483

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

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

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

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

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

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.

Table 1. Key manipulations across the six experiments. Stimuli for the two talkers was interleaved during exposure for all experiments. Across experiments, we manipulated whether gender of the two talkers was different or was the same. Within each of the talker gender manipulations, we manipulated whether test was blocked by talker or interleaved across talkers, and whether talker acoustics varied across talkers or were identical across talkers.

Experiment	Exposure	Talker gender	Test	Talker acoustics
1	Interleaved	Different	Blocked	Vary across talkers
2	Interleaved	Different	Interleaved	Vary across talkers
3	Interleaved	Different	Interleaved	Identical across talkers
4	Interleaved	Same	Blocked	Vary across talkers
5	Interleaved	Same	Interleaved	Vary across talkers
6	Interleaved	Same	Interleaved	Identical across talkers

Table 2. Results of the generalized linear mixed effects models examining learning within each dose condition for each of the three different gender experiments (experiments 1 – 3).

Experiment	**Dose**	**Fixed effect**	$\widehat{\beta}$	***SE***	*z*	*p*
1	1x	(Intercept)	-2.847	0.389	-7.323	< 0.001
		Step	4.876	0.393	12.409	< 0.001
		Bias	2.218	0.501	4.428	< 0.001
		Step x Bias	-1.430	0.623	-2.293	0.022
	2x	(Intercept)	-2.550	0.261	-9.774	< 0.001
		Step	4.831	0.317	15.217	< 0.001
		Bias	1.463	0.609	2.402	0.016
		Step x Bias	-0.962	0.644	-1.495	0.135
2	1x	(Intercept)	-2.492	0.309	-8.050	< 0.001
		Step	4.979	0.401	12.420	< 0.001
		Bias	2.246	0.529	4.243	< 0.001
		Step x Bias	-1.136	0.628	-1.809	0.070
	2x	(Intercept)	-2.184	0.256	-8.544	< 0.001
		Step	4.400	0.303	14.543	< 0.001
		Bias	1.944	0.483	4.020	< 0.001
		Step x Bias	-0.176	0.541	-0.325	0.745
3	1x	(Intercept)	-1.129	0.246	-4.585	< 0.001
		Step	3.148	0.229	13.755	< 0.001
		Bias	1.022	0.540	1.892	0.058
		Step x Bias	-0.849	0.413	-2.056	0.040
	2x	(Intercept)	-1.276	0.269	-4.740	< 0.001
		Step	3.321	0.221	15.038	< 0.001
		Bias	1.243	0.626	1.987	0.047
		Step x Bias	-0.340	0.483	-0.704	0.481

Table 3. Results of the generalized linear mixed effects model examining learning between dose conditions for each of the three different gender experiments (experiments 1 – 3).

Experiment	**Fixed effect**	$\widehat{\beta}$	***SE***	***z***	***p***
1	(Intercept)	-2.692	0.231	-11.632	< 0.001
	Step	4.851	0.248	19.537	< 0.001
	Bias	1.802	0.391	4.613	< 0.001
	Dose	0.202	0.418	0.484	0.628
	Step x Bias	-1.137	0.443	-2.570	0.010
	Step x Dose	0.088	0.405	0.216	0.829
	Bias x Dose	-1.061	0.665	-1.595	0.111
	Step x Bias x Dose	1.126	0.652	1.727	0.084
2	(Intercept)	-2.311	0.196	-11.771	< 0.001
	Step	4.641	0.241	19.265	< 0.001
	Bias	2.031	0.351	5.793	< 0.001
	Dose	0.154	0.356	0.434	0.665
	Step x Bias	-0.551	0.400	-1.375	0.169
	Step x Dose	-0.231	0.404	-0.571	0.568
	Bias x Dose	0.092	0.614	0.150	0.881
	Step x Bias x Dose	0.339	0.588	0.576	0.564
3	(Intercept)	-1.198	0.184	-6.526	< 0.001
	Step	3.227	0.162	19.869	< 0.001
	Bias	1.121	0.415	2.697	0.007
	Dose	-0.128	0.348	-0.366	0.714
	Step x Bias	-0.591	0.320	-1.846	0.065
	Step x Dose	0.143	0.288	0.496	0.620
	Bias x Dose	0.408	0.799	0.511	0.610
	Step x Bias x Dose	-0.027	0.558	-0.048	0.962

Table 4. Results of the generalized linear mixed effects models examining learning within each dose condition for each of the three same gender experiments (experiments 4 – 6).

Experiment	Dose	Fixed effect	$\hat{\beta}$	*SE*	*z*	*p*
4	1x	(Intercept)	-2.618	0.251	-10.450	< 0.001
		Step	5.162	0.382	13.510	< 0.001
		Bias	0.198	0.446	0.443	0.658
		Step x Bias	0.983	0.770	1.275	0.202
	2x	(Intercept)	-2.151	0.287	-7.491	< 0.001
		Step	4.612	0.353	13.053	< 0.001
		Bias	0.770	0.420	1.832	0.067
		Step x Bias	-0.397	0.490	-0.810	0.418
5	1x	(Intercept)	-1.640	0.251	-6.542	< 0.001
		Step	4.179	0.310	13.497	< 0.001
		Bias	1.442	0.428	3.369	0.001
		Step x Bias	-0.669	0.431	-1.552	0.121
	2x	(Intercept)	-1.812	0.301	-6.015	< 0.001
		Step	4.148	0.296	14.013	< 0.001
		Bias	1.041	0.434	2.399	0.016
		Step x Bias	-0.368	0.427	-0.861	0.389
6	1x	(Intercept)	-0.932	0.218	-4.277	< 0.001
		Step	3.123	0.241	12.976	< 0.001
		Bias	0.319	0.405	0.787	0.431
		Step x Bias	-0.364	0.350	-1.041	0.298
	2x	(Intercept)	-1.085	0.239	-4.541	< 0.001
		Step	3.132	0.244	12.852	< 0.001
		Bias	0.904	0.451	2.005	0.045
		Step x Bias	-0.497	0.432	-1.151	0.250

Table 5. Results of the generalized linear mixed effects model examining learning between dose conditions for each of the three same gender experiments (experiments 4 – 6).

Experiment	Fixed effect	$\widehat{\beta}$	*SE*	*z*	*p*
4	(Intercept)	-2.393	0.194	-12.363	< 0.001
	Step	4.882	0.259	18.823	< 0.001
	Bias	0.502	0.305	1.643	0.100
	Dose	0.475	0.355	1.341	0.180
	Step x Bias	0.233	0.447	0.522	0.602
	Step x Dose	-0.440	0.433	-1.016	0.310
	Bias x Dose	0.419	0.520	0.806	0.420
	Step x Bias x Dose	-0.754	0.668	-1.128	0.259
5	(Intercept)	-1.722	0.192	-8.991	< 0.001
	Step	4.165	0.216	19.273	< 0.001
	Bias	1.221	0.300	4.072	< 0.001
	Dose	-0.199	0.365	-0.544	0.586
	Step x Bias	-0.496	0.299	-1.656	0.098
	Step x Dose	-0.065	0.384	-0.170	0.865
	Bias x Dose	0.148	0.533	0.277	0.782
	Step x Bias x Dose	-0.391	0.390	-1.003	0.316
6	(Intercept)	-0.999	0.161	-6.207	< 0.001
	Step	3.117	0.169	18.407	< 0.001
	Bias	0.592	0.300	1.973	0.048
	Dose	-0.112	0.312	-0.358	0.720
	Step x Bias	-0.417	0.272	-1.534	0.125
	Step x Dose	-0.093	0.313	-0.296	0.767
	Bias x Dose	0.267	0.580	0.460	0.646
	Step x Bias x Dose	0.246	0.467	0.527	0.598

Table 6. Results of the generalized linear mixed effects model for reaction time at test.

Fixed effect	$\widehat{\boldsymbol{\beta}}$	***SE***	***z***	***p***
(Intercept)	629.773	12.608	49.951	0.000
Gender	22.816	14.440	1.580	0.114
E1/E4 – E2/E5	48.160	17.623	2.733	0.006
E2/E5 – E3/E6	56.881	17.741	3.206	0.001
Dose	-1.920	14.440	-0.133	0.894
Gender x E1/E4 – E2/E5	44.621	35.247	1.266	0.206
Gender x E2/E5 – E3/E6	-7.648	35.482	-0.216	0.829
Gender x Dose	-14.628	28.879	-0.507	0.612
E1/E4 – E2/E5 x Dose	14.553	35.247	0.413	0.680
E2/E5 – E3/E6 x Dose	-33.751	35.482	-0.951	0.341
Gender x E1/E4 – E2/E5 x Dose	-93.591	70.494	-1.328	0.184
Gender x E2/E5 – E3/E6 x Dose	-19.972	70.965	-0.281	0.778

Figure 1. Acoustic characteristics of the stimuli used in the different gender experiments (experiments 1 – 3). Points in black connected by a line indicate test tokens; all other points indicate exposure tokens. As described in the main text, fundamental frequency was measured for the voiced portion of each token and center of gravity was measured for the fricative portion of each token.

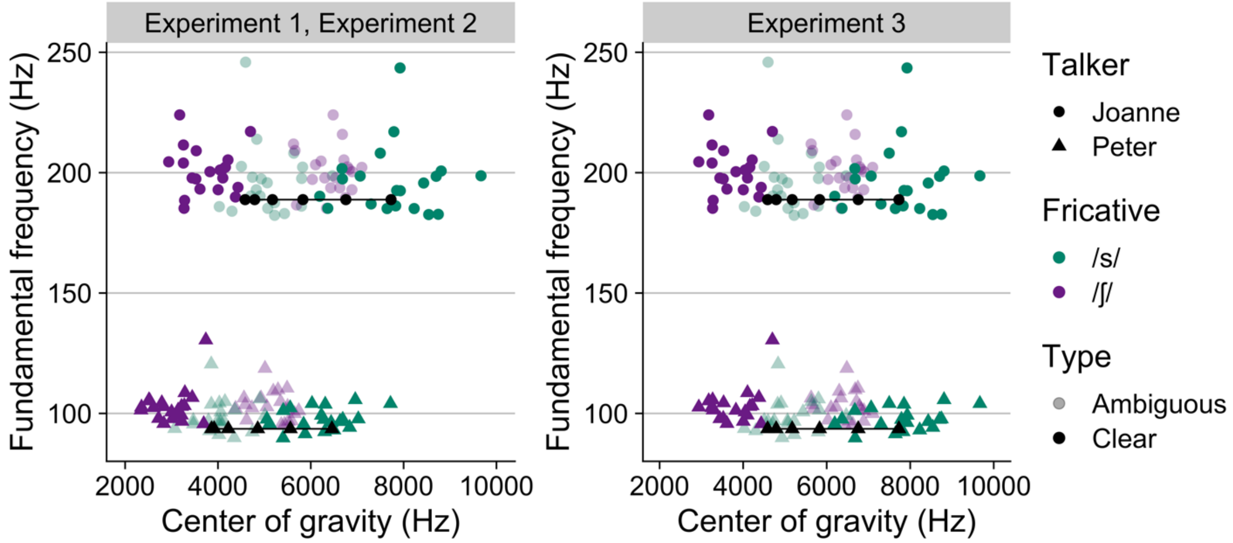
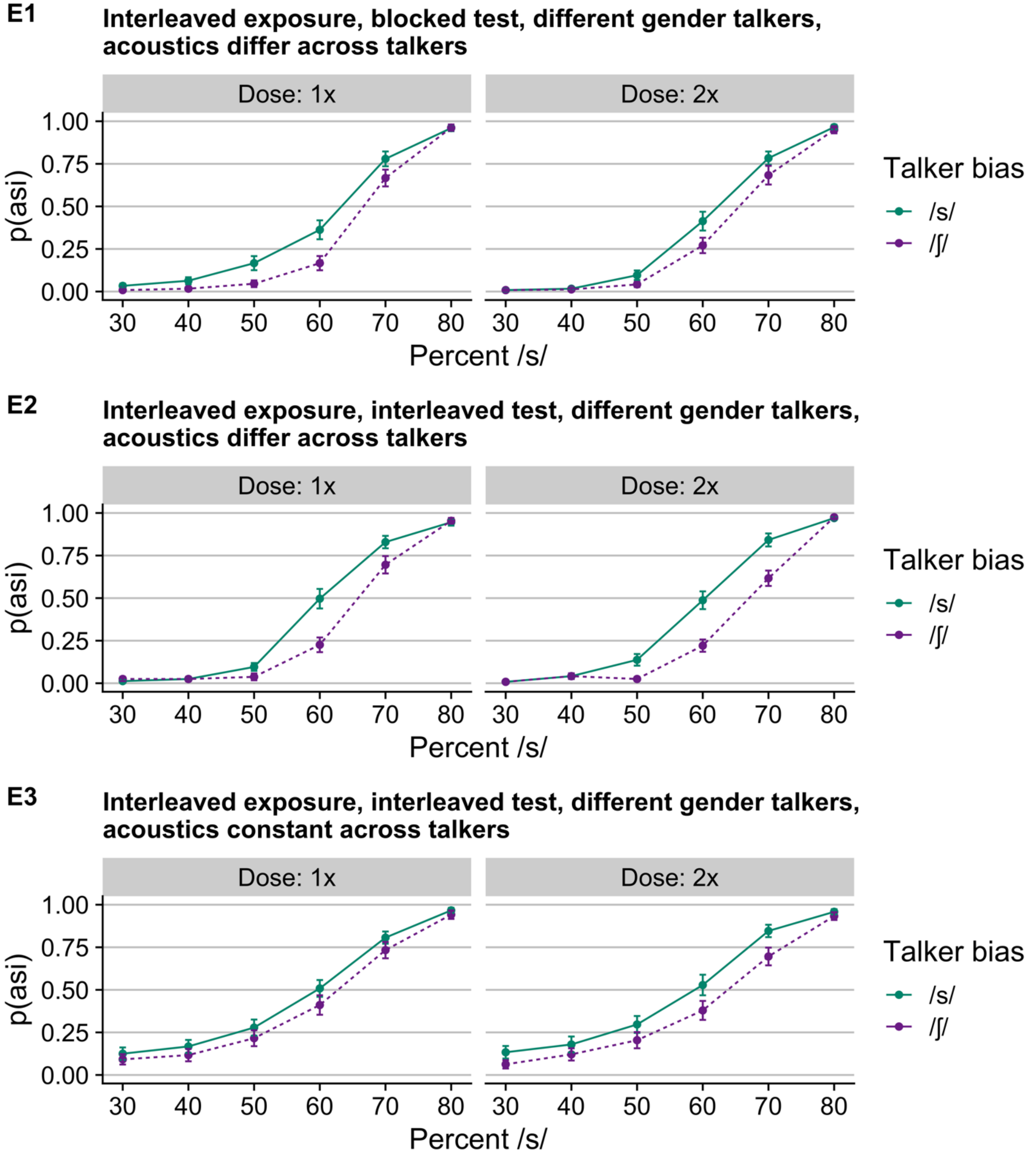


Figure 2. Mean proportion *asi* responses as a function of continuum step, talker bias, and dose for the different gender experiments, which included experiment 1 (E1), experiment 2 (E2), and experiment 3 (E3). As described in the main text, talker bias was manipulated within subjects, with exposure stimuli selected to differentially bias listeners to perceive ambiguous variants as /s/ for one talker and /ʃ/ for the other talker. Dose was manipulated between subjects. Continuum step is presented in terms of percent /s/ energy in each step of the test continuum. Means reflect grand means calculated over by-subject means; error bars indicate standard error of the mean.



E1 Interleaved exposure, blocked test, different gender talkers, acoustics differ across talkers

E2 Interleaved exposure, interleaved test, different gender talkers, acoustics differ across talkers

E3 Interleaved exposure, interleaved test, different gender talkers, acoustics constant across talkers

Figure 3. Acoustic characteristics of the stimuli used in the same gender experiments (experiments 4 – 6). Points in black connected by a line indicate test tokens; all other points indicate exposure tokens. As described in the main text, fundamental frequency was measured for the voiced portion of each token and center of gravity was measured for the fricative portion of each token.

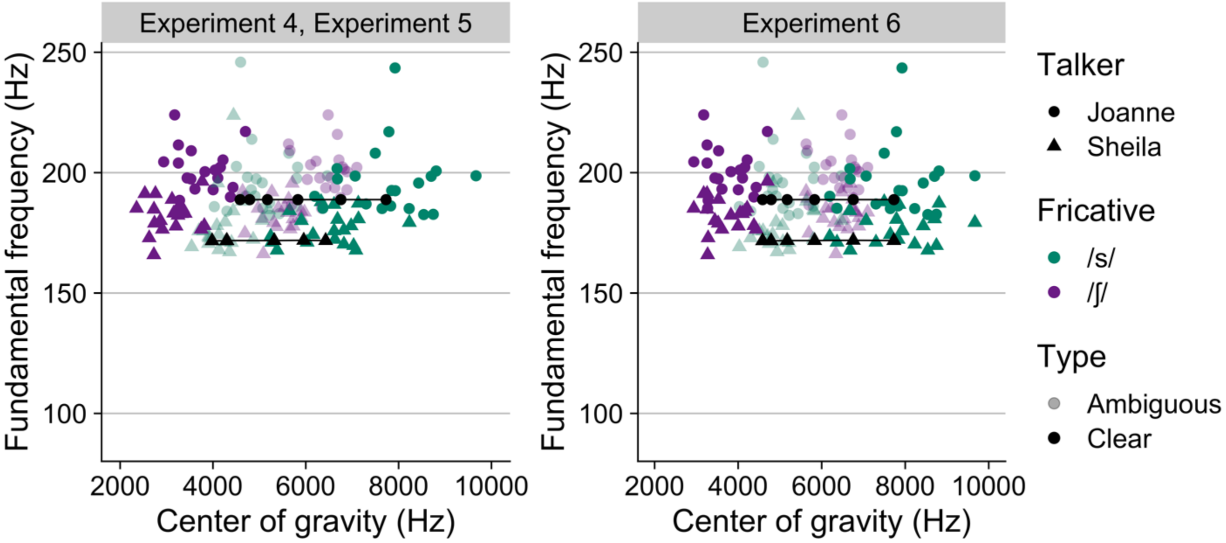


Figure 4. Mean proportion *asi* responses as a function of continuum step, talker bias, and dose for the same gender experiments, which included experiment 4 (E4), experiment 5 (E5), and experiment 6 (E6). As described in the main text, talker bias was manipulated within subjects, with exposure stimuli selected to differentially bias listeners to perceive ambiguous variants as /s/ for one talker and /ʃ/ for the other talker. Dose was manipulated between subjects. Continuum step is presented in terms of percent /s/ energy in each step of the test continuum. Means reflect grand means calculated over by-subject means; error bars indicate standard error of the mean.

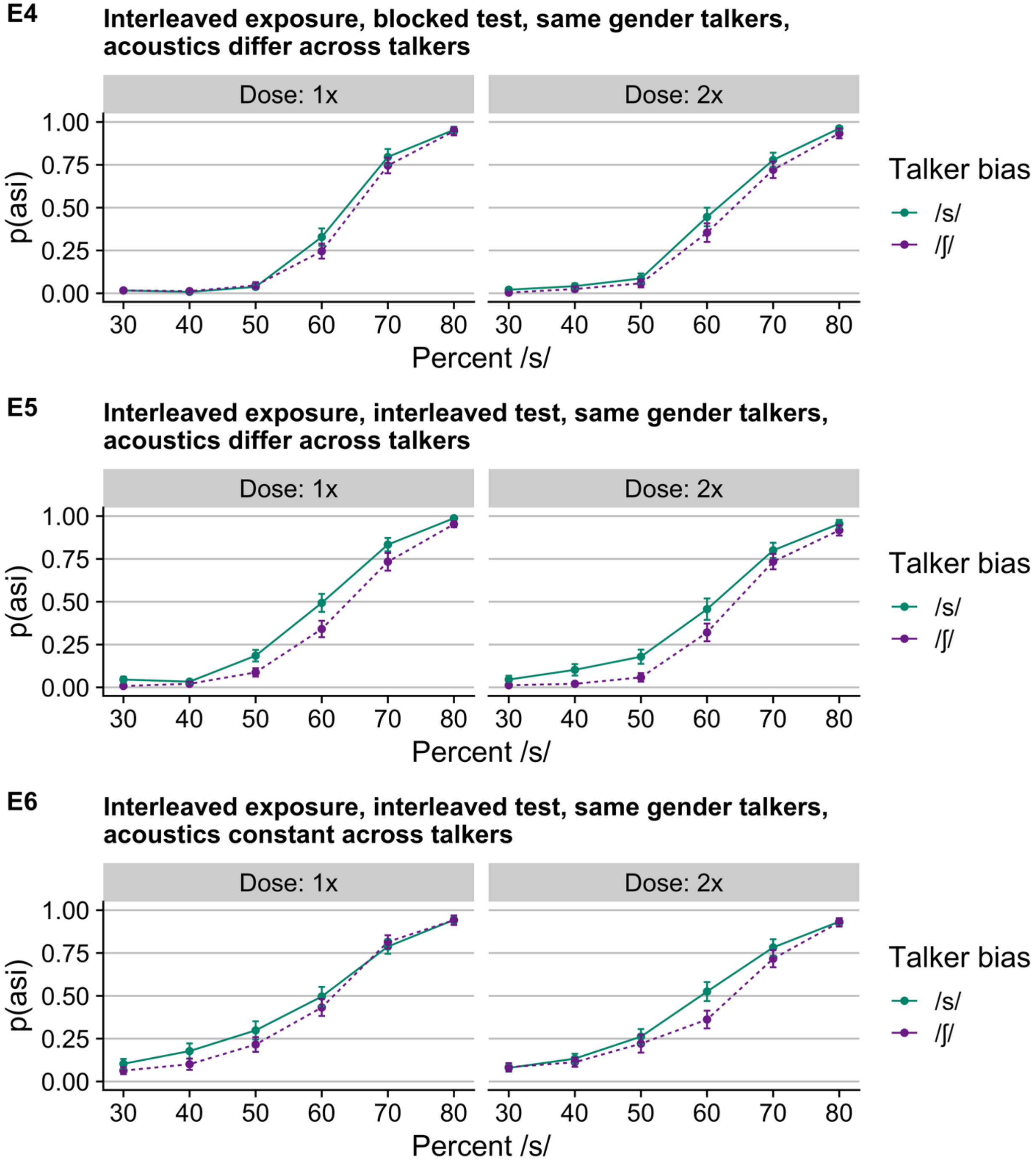


Figure 5. Learning effect sizes across the experiments reported here and those in Luthra et al. (2021, labeled LMM in this figure). Effect size is represented by the beta estimate for the fixed effect of bias in each regression model; error bars show the 95% confidence interval. A beta estimate of zero corresponds to no learning. The region shown in gray reflects the 95% confidence interval of the learning effect size in experiment 1A of Tzeng et al. (2021), which used the same stimuli as the f1 talker in the current work and the standard learning paradigm in which listeners only heard one talker during exposure (and test), with bias manipulated between-subjects.

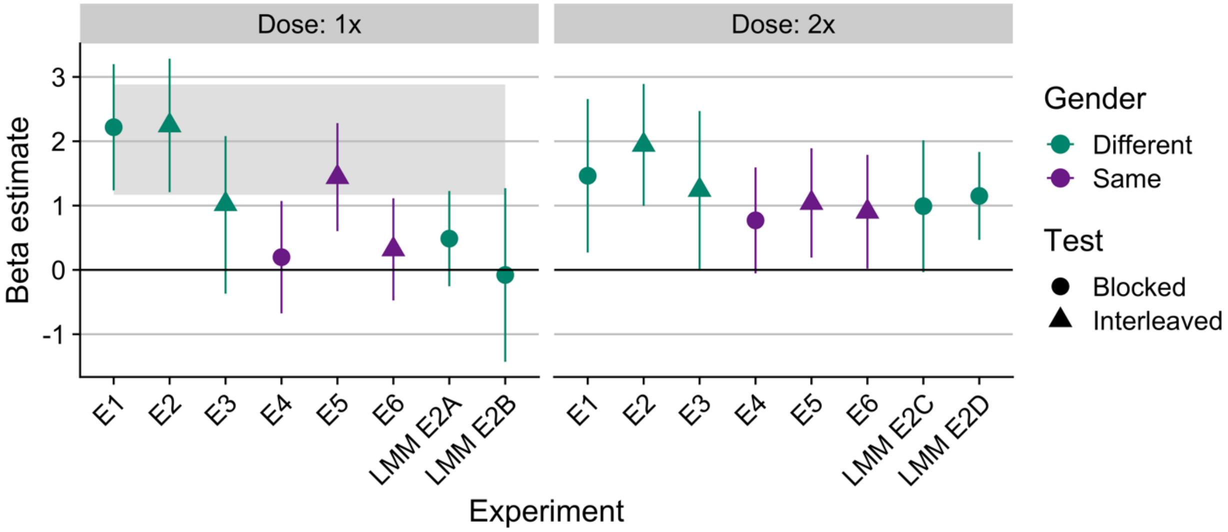
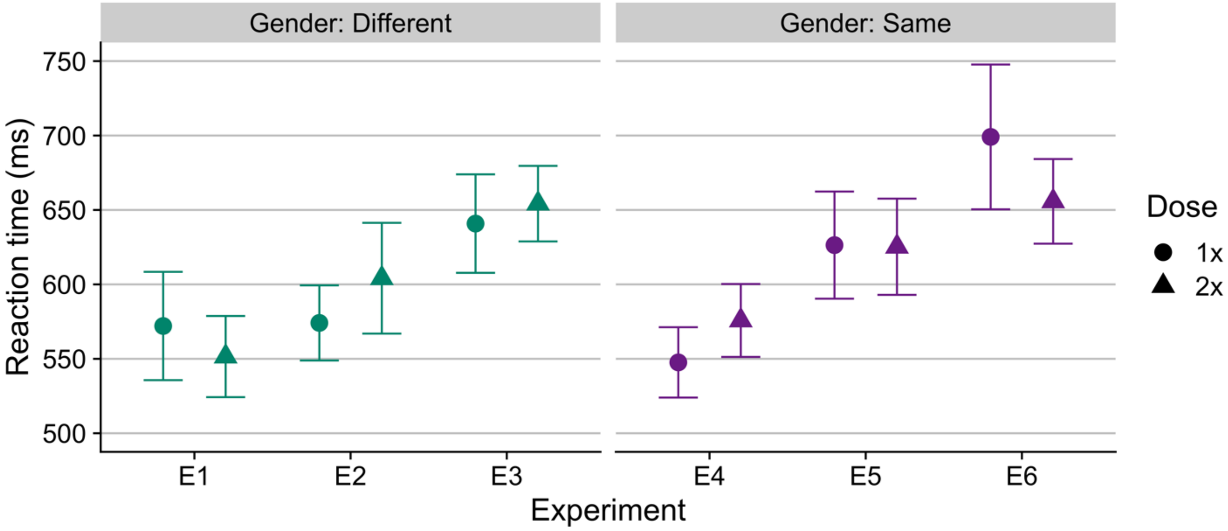


Figure 6. Mean reaction time in milliseconds (ms) for each experiment and each dose condition. Error bars indicate standard error of the mean.


Gender: Different
Gender: Same
Reaction time (ms)
E1
E2
E3
E4
E5
E6
Experiment
Dose
● 1x
▲ 2x