

# Talker-Specific Pronunciation or Speech Error? Discounting (or not) Atypical Pronunciations During Speech Perception

Linda Liu and T. Florian Jaeger  
University of Rochester

Perceptual recalibration allows listeners to adapt to talker-specific pronunciations, such as atypical realizations of specific sounds. Such recalibration can facilitate robust speech recognition. However, indiscriminate recalibration following any atypically pronounced words also risks interpreting pronunciations as characteristic of a talker that are in reality because of incidental, short-lived factors (such as a speech error). We investigate whether the mechanisms underlying perceptual recalibration involve inferences about the causes for unexpected pronunciations. In 5 experiments, we ask whether perceptual recalibration is blocked if the atypical pronunciations of an unfamiliar talker can also be attributed to other incidental causes. We investigated 3 types of incidental causes for atypical pronunciations: the talker is intoxicated, the talker speaks unusually fast, or the atypical pronunciations occur only in the context of tongue twisters. In all 5 experiments, we find robust evidence for perceptual recalibration, but little evidence that the presence of incidental causes block perceptual recalibration. We discuss these results in light of other recent findings that incidental causes can block perceptual recalibration.

## Public Significance Statement

This study investigates the mechanisms operating during human speech perception. The results suggest limits in the types of information that can be integrated during real-time processing of spoken language.


*Keywords:* speech perception, perceptual recalibration, causal inference, speech error, tongue twister

Talkers differ from each other in their meaning-to-sound mappings: the same word produced in the same context will differ acoustically and phonetically depending on the talker. How listeners overcome this problem has continued to be one of the pressing questions in research on speech perception. One important part of the answer seems to be adaptive mechanisms during speech perception. Adaptation is observed when listeners are exposed to unfamiliar talkers with nonnative or otherwise atypical pronunciations (e.g., Bradlow & Bent, 2008; Sidas, Alexander, &

Nygaard, 2009; Xie, Theodore, & Myers, 2017). While listeners might initially experience processing difficulty, some of this difficulty can be overcome within minutes of exposure (Clarke & Garrett, 2004; Xie et al., 2018). The adaptive nature of the speech perception system is also evident in a phenomenon called perceptual recalibration. When exposed to an unfamiliar talker with atypical pronunciations of a sound category, listeners adapt the categorization boundary between those sound categories (e.g., Eisner & McQueen, 2006; Kraljic & Samuel, 2005; Norris, McQueen, & Cutler, 2003; Reinisch & Holt, 2014; Vroomen & Baart, 2009). For example, after exposure to a talker who produces /s/ in a way that makes it sound more like an /ʃ/, listeners change the boundary along the /s-/ʃ/ continuum, so that more sounds along that continuum are now categorized as /s/.<sup>1</sup>

Intuitively, recalibration facilitates robust speech perception, helping listeners to overcome intertalker variability in the sound-meaning mapping. While the existence of perceptual recalibration is now firmly established, questions remain about the nature of its underlying mechanisms (for review, see Weatherholtz & Jaeger, 2016). Here we ask whether recalibration applies indiscriminately when an unfamiliar talker with atypical pronunciation is encountered, or whether perceptual recalibration can be cancelled if there

---

Linda Liu, Department of Brain and Cognitive Sciences, University of Rochester;  T. Florian Jaeger, Department of Brain and Cognitive Sciences and Department of Computer Science, University of Rochester.

The research presented here was funded by NIH R01 Grant HD075797 to T. Florian Jaeger. The views expressed here do not necessarily reflect those of the funding agency. The authors are grateful for particularly helpful feedback from John Alderete, Athanassios Protopapas, Arty Samuel, Rachel Theodore, and Xin Xie. Earlier presentations of this work benefitted from feedback from Ehsan Hogue, Crystal Lee, Michael K. Tanenhaus, Davy Temperley, Xin Xie, as well as members of the Human Language Processing lab at the University of Rochester. Additional online data are available at <https://osf.io/ungba/>.

Correspondence concerning this article should be addressed to T. Florian Jaeger, Department of Brain and Cognitive Sciences, University of Rochester, Meliora Hall, Box 270268, Rochester, NY 14627-0268. E-mail: [fjaeger@ur.rochester.edu](mailto:fjaeger@ur.rochester.edu)

---

<sup>1</sup> Throughout this article, slashes indicate phonological transcriptions based on the international phonetic alphabet. The symbol /ʃ/ refers to the sound spelled “sh” in English.

is evidence that the input is not characteristic of the talker—for example, because the pronunciation might have resulted from an *incidental* cause (e.g., the talker is chewing gum).

In an influential study, Kraljic, Samuel, and Brennan (2008) found that perceptual recalibration to atypical pronunciations of /f/ was blocked when the atypical pronunciations could be attributed to an incidental cause. In their experiments, atypical pronunciations were either paired with a video showing the talker producing the shifted word with a pen in their mouth or with a pen in their hand. Kraljic and colleagues identified perceptual recalibration when the shifted pronunciations were paired with videos where the talker has a pen in the hand. When the talker had a pen in the mouth while producing the atypical sound, perceptual recalibration was blocked. One explanation for this blocking is that listeners attribute the atypical pronunciations to the pen (Liu & Jaeger, 2018; for related discussion, see Arnold, Kam, & Tanenhaus, 2007; Kraljic et al., 2008). Inferences about the causes for unexpected pronunciations would allow listeners to determine whether they should expect the same talker to sound similar on future occasions, or whether the observed deviation from expected pronunciations was incidental (though alternative explanations have been proposed; Kraljic & Samuel, 2011).

As of yet, this *pen-in-the-mouth* effect remains the only manipulation of incidental causes for which blocking of perceptual recalibration has been investigated. Thus, it is an open question as to whether other incidental causes can block (or at least reduce) recalibration, as would be expected if causal inferences underlie the pen-in-the-mouth effect. More generally, relatively little is known about the extent to which listeners take into account alternative causes when interpreting linguistic input. Some studies have found similar effects on other aspects of language understanding—in particular, alternative causes presented in explicit instructions (e.g., Arnold et al., 2007; Dix et al., 2018; Grodner & Sedivy, 2011; Kurumada, Brown, Bibyk, & Tanenhaus, 2018, as summarized in Rohde & Kurumada, 2018). For example, listeners tend to anticipate unfamiliar objects as referents after a speech disfluency (“Click on [pause] thee uh red . . .”), as evidenced in anticipatory eye-movements in a visual world paradigm (Arnold et al., 2007). This effect was blocked when listeners were told that the speaker suffered from a pathology that made naming objects difficult. Results like these suggest that listeners can in principle integrate the presence of alternative causes for the linguistic input they observe during the interpretation of that input. Whether similar inferences can affect perceptual recalibration or other adaptive processes during speech perception remains an open question.

Here we investigate listeners’ perceptual recalibration when atypical pronunciations of /s/ or /f/ are presented in the context of incidental causes. Across five experiments, we investigate the effects of three incidental causes: alleged intoxication, faster than usual speech rate, and tongue twisters. Any of these factors can cause atypical pronunciation of the /s-/f/ contrast, though our focus lies on tongue twisters. Anyone who grew up in an English-speaking environment is likely familiar with well-known tongue twisters like “She sells seashells by the seashore.” or “Peter Piper picked a peck of pickled peppers.” Tongue twisters are notoriously difficult for talkers to produce and often result in speech errors when produced quickly. This is precisely the property that makes tongue twisters a suitable manipulation for the present purpose. While categorical speech errors—such as full phoneme exchanges—

are rare in spontaneous speech (<0.1–2%; as estimated in Garnham, Shillcock, Brown, Mill, & Cutler, 1981; Levelt, 1993; Wijnen, 1992), the rate of speech errors increases drastically when speakers have to produce sequences of similar sounding words in production experiments (up to 8–17%; according to Choe & Redford, 2012; Motley & Baars, 1976) and even more so in the context of tongue twisters. Because of perceptual biases, these numbers likely underestimate the true rate of speech errors (by some estimates by a factor of three or more; Alderete & Davies, 2018; Ferber, 1991). We draw on this increased incidence of production errors in tongue twister contexts compared with nontongue twister contexts. Specifically, we ask whether participants are less likely to expect all pronunciations of a talker to sound atypical when all previously observed atypical pronunciations by that talker occurred in tongue twisters, compared with when previously observed atypical pronunciations occurred in nontongue twister contexts.

All our experiments use graded phonetic deviations from typical pronunciations, rather than categorical speech errors. Traditionally, the study of speech errors in productions has focused on categorical errors (phoneme substitution, deletion, transposition, omission, or addition; e.g., Fromkin, 1971). However, recent analyses suggest that speech errors are often graded noncategorical deviations from the intended pronunciation (Frisch & Wright, 2002; Goldrick & Blumstein, 2006; Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007; McMillan & Corley, 2010; Mowrey & MacKay, 1990; Pouplier, 2007). For example, Frisch and Wright (2002) measured the percentage of voicing, duration of frication, and the amplitude of frication for the /s/ and /z/ contrasts, produced in a tongue twister context. Frisch and Wright found that errors often exhibited phonetic characteristics that placed them along the continuum between /s/ and /z/, rather than being categorical substitution of one sound for the other. Similarly, Navas (2001, as cited in Goldrick & Blumstein, 2006) found that some fricative errors exhibited spectral characteristics that were between typical /s/ and /ʃ/ pronunciations (for a concise summary of related works; see Alderete & Davies, 2018, pp. 27–29).

This suggests that tongue twisters are a suitable incidental cause for the present purpose: similar (experimenter-created) gradient pronunciations are used in perceptual recalibration experiments, including the experiments we present here. Imagine you hear a talker produce the tongue twister “She sells seashells by the seashore.” The talker might pronounce the beginning of this phrase as “She shells,” shifting the /s/ in “sells” toward (but not completely) the /ʃ/ in “shells.” If listeners can take into account incidental causes, they should infer that this pronunciation might not be typical for the talker and, therefore, not predictive of future pronunciations of /s/ by the same talker. We would expect that perceptual recalibration is reduced if the talker’s shifted pronunciations only ever occur in the context of tongue twisters.

## Overview of Experiments

Experiment 1 verifies that our paradigm can detect perceptual recalibration. Finding this confirmed, we test whether perceptual recalibration is blocked when shifted sounds during exposure only occur within a tongue twister context (e.g., “passion mansion passive passion”). Blocking of perceptual recalibration is expected if listeners’ fully attribute the atypical pronunciation to the tongue

twister context and, thus, infer that those atypical pronunciations are *not* informative about how the same talker's speech outside of tongue twister contexts.

Anticipating our results, we find robust evidence of perceptual recalibration. However, we do not find significant blocking of perceptual recalibration in the tongue twister condition: the perceptual recalibration effect in Experiment 1 does not differ significantly across nontongue twister and tongue twister contexts. This leads us to conduct Experiment 2, which establishes that we can in principle detect statistically significant differences between exposure that elicits perceptual recalibration (as in Experiment 1) and exposure that does not elicit perceptual recalibration (as in Experiment 2). Experiment 3 explores whether the presentation of explicit instructions about plausible incidental causes for shifted pronunciations—for example, that the talker is intoxicated—can block perceptual recalibration. We again find robust perceptual recalibration effects across all conditions, and no significant evidence that incidental causes can block perceptual recalibration. This leads us to assess the plausibility of our tongue twister contexts, and compare them against attested tongue twisters like “Peter Piper Pepper Peter.” Experiment 4 identifies the most convincing tongue twister contexts and assesses whether perceptual recalibration can be blocked when only those most plausible tongue twisters are used. We again observe robust perceptual recalibration after exposure to nontongue twister contexts. Again, we find no significant blocking of perceptual recalibration after exposure to tongue twister contexts. Finally, Experiment 5 tests whether perceptual recalibration is reduced if the shifted pronunciation occurs together with clear signs of production difficulty.

Like the planned analyses for Experiments 1–4, Experiment 5 fails to find significant evidence that listeners integrate incidental causes to explain away atypical pronunciations. These findings contrast with the robust pen-in-the-mouth effect, which has been replicated across a number of experiments (Kraljic & Samuel, 2011; Kraljic et al., 2008), including in paradigms similar to the one used here (Liu & Jaeger, 2018). There is, however, *some* evidence in support of causal inference during perceptual recalibration: the nonsignificant effects we observe go in the predicted direction (reduced perceptual recalibration in the presence of an incidental cause) in five out of six between-subjects comparisons. Prompted by reviewers, we conducted post hoc analyses. These analyses reveal some (albeit weak) evidence consistent with the hypothesis that incidental causes can reduce the magnitude of perceptual recalibration.

In the general discussion, we review how our results narrow down possible explanations for the effect of visually presented causes like the pen in the mouth. Broadly speaking, one possibility is that the pen-in-the-mouth effect does *not* originate in causal inferences, contrary to our earlier interpretation (Liu & Jaeger, 2018). This would, however, raise the need for alternative explanations of previous findings that have been attributed to causal inferences (for discussion, see Kraljic & Samuel, 2011). Another possibility is that the pen-in-the-mouth effect *does* originate in causal inferences but that visual information, or specifically visual information about articulation, has a special status during speech processing—for example, because of special mechanisms dedicated to the integration of audio-visual percepts (cf. Rosenblum, 2008; Tuomainen, Andersen, Tiippana, & Sams, 2005). Finally, our results are compatible with the hypothesis that perceptual

recalibration is affected by causal inferences, *provided that these inferences are exquisitely sensitive to the probability of the hypothesized incidental cause resulting in the observed auditory percepts*. We discuss the properties of our experiments that afford this latter interpretation, and determine future steps to distinguish between the different accounts.

## Analysis and Reporting Approach

Following standard procedure from our lab, we report all studies conducted for this project. Three auxiliary experiments that yielded identical results to Experiments 1–5 are presented in online supplemental information available via OSF <https://osf.io/ungba/>, and summarized in the main text. Unless explicitly mentioned otherwise, analyses were planned before any inspection of the data.

The number of participants and test items included in the analysis was held constant across all experiments, and was chosen so as to achieve sufficient statistical power based on the effect sizes reported in similar previous research (for details, see Method). We confirmed that we have high power by parametrically generating 10,000 data sets with an effect size estimated from previous work—specifically, half the estimate observed in Liu and Jaeger (2018). These estimates were intended, and turn out, to be conservative (the effect sizes observed in the experiments reported below are *larger* than those assumed in the power analyses). Simulations estimate our power to detect perceptual recalibration (Label effect in predicted direction) at >95%, and the power to detect blocking of perceptual recalibration (interaction of Label and Context effects in predicted direction) at >81% (for explanation of the conditions, see below). All data and analyses are available at <https://osf.io/ungba/>.

## Aggregate Demographic Information About Participants

Because the demographic composition of our participants did not vary significantly across experiments, we report aggregate information here. All demographic categories were based verbatim on National Institutes of Health (NIH) reporting requirements. Across all experiments presented here, 48% of our participants reported as female, and 47% report as male, and 5% declined to report gender. The mean age of our participants was 36.3 years, with an interquartile range of 27–42 years ( $SD = 19$ ; 4% declined to report). All participants reported to be at least 18 years of age. With regard to ethnicity, 9% of the participants reported as Hispanic, 85% as Non-Hispanic, and 6% declined to report. With regard to race, 74% report as White, 8% as Black or African American, 7% as Asian, 4% as More than one race, 1% as American Indian/Alaska Native or Native Hawaiian or other Pacific Islander, 1% as other, and 5% declined to report. As we have no theoretical reasons to investigate demographic effects on the outcomes reported in the present study, we refrained from doing so.

## Experiment 1

We begin by verifying that we can detect perceptual recalibration to atypical pronunciations of /s/ and /ʃ/ in a new variant of an



exposure-test paradigm that accommodates our present goals. The general structure of our experiments is summarized in Figure 1 and elaborated on below. Following previous perceptual recalibration experiments, our experiment consisted of an exposure block intended to induce perceptual recalibration, followed by a test block to assess the degree of perceptual recalibration (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2011; Kraljic et al., 2008; Liu & Jaeger, 2018; Norris et al., 2003).

To study the effects of tongue twisters, listeners in the present study heard four word phrases during exposure, some of which contained the shifted /ʔsʃ/ sound (either an /s/ shifted toward and /f/ or vice versa; for details, see Method). Specifically, we used a 2 × 2 between-participants design in the exposure block (Label × Context). Participants heard a shifted sound replace the /s/ sound (S-Label condition) or the /f/ sound (f-Label condition), and these atypical pronunciations either occurred in a Tongue Twister Context (e.g., “passive massive paʔsʃion passive”) or a Non-Tongue Twister Context (e.g., “holler tamper paʔsʃion holler”). The Tongue Twister Context contained sound sequences intended to make it more difficult to produce than the Non-Tongue Twister Context. This was intended to make it seem likely to participants that any atypical pronunciation in the Tongue Twister Context was because of an incidental speech error. We use the /s/ and /f/ contrast because these sounds are commonly exchanged for each other in speech errors (Shattuck-Hufnagel & Klatt, 1979). This makes it more likely that atypical pronunciations of /s/ and /f/ in a tongue twister context will be seen as a plausible speech error.

The test block did not vary across participants, and followed previous perceptual recalibration experiments (Kraljic & Samuel, 2005; Liu & Jaeger, 2018; Norris et al., 2003). During test, we assess whether exposure affected categorization along an /s/-/f/ continuum, as expected from previous studies on perceptual recalibration. We then examine whether this perceptual recalibration effect could be blocked or reduced depending on the context in which /s/ and /f/ appeared.

All experiments reported below use a web-based crowdsourcing paradigm. This allows us to collect data more quickly, and from a more heterogeneous participant group than lab-based paradigms. This was particularly helpful for the present studies, which include a total of 960 participants. We have used similar web-based paradigms in previous work on speech perception (e.g., Bicknell, Bushong, Tanenhaus, & Jaeger, 2019; Burchill, Liu, & Jaeger, 2018; Bushong & Jaeger, 2017; Kleinschmidt, Raizada, & Jaeger, 2015; Xie et al., 2018), including lexically guided perceptual recalibration to /s/ and /f/ (Liu & Jaeger, 2018).

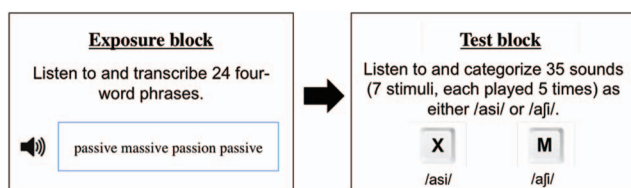


Figure 1. Structure of experiments. During the exposure block, participants heard 24 four-word phrases and were asked to transcribe them. Exposure was manipulated between participants. During the test block, all participants categorized sounds as either /asi/ or /afj/. See the online article for the color version of this figure.

## Method

**Participants.** There were 173 total participants recruited to achieve a target of 40 participants for each of the four between-participants conditions (S/f-Label crossed with the Tongue Twister/Non-Tongue Twister Context). The same holds for all perceptual recalibration experiments presented below, to avoid unnecessary researchers’ degrees of freedom. The targeted number of participants is comparable with previous experiments on perceptual recalibration (e.g., ~48 in Kraljic & Samuel, 2005; ~25 participants Norris et al., 2003).

The experiment took about 10 min, and participants were paid \$1.00 (\$6/hr). Participants were instructed to participate only if they were native speakers of English, and if they would complete the experiment while wearing headphones in a quiet room.

Exclusion criteria were determined before conducting the experiment, closely following previous work (specifically, all applicable criteria from Liu & Jaeger, 2018). Based on these criteria, eight participants were excluded based on their incorrect response to a catch question asking them to identify whether the exposure talker was male or female (the talker was clearly a female speaker), and two participants were excluded for reporting that they did not wear headphones during the experiment (7.5% total exclusion rate). Both the catch question and the headphone question were part of a postexperiment exit survey described below. Additionally, three participants were excluded for likely confusing their response keys during the test block, as evidenced by inverted categorization boundaries (more /s/ responses at the /f/ end of the continuum), which would not be expected under any theory of speech perception.

### Materials.

**Exposure block: Transcription task.** Participants heard and transcribed 24 four-word phrases (all phrases given in Tables 1 and 2). The condition (Label × Context) that the participant was in dictated the specific set of phrases they would hear. 8 of these phrases contained a shifted pronunciation of either /s/ or /f/, depending on the Label condition. We chose to have eight shifted pronunciations because we had previously found perceptual recalibration to /s/ and /f/ in similar paradigms with six and 10 shifted pronunciations, and little to no benefit for more than 10 shifted pronunciations (Liu & Jaeger, 2018; see also Kleinschmidt & Jaeger, 2011). Our power simulations were based on half the effect size found in previous work for 10 shifted pronunciations (see Appendix for details).

We refer to the phrases that contained a shifted pronunciation as critical phrases. The eight critical phrases occurred in either a Tongue Twister or Non-Tongue Twister Context, described below. The other 16 phrases were filler phrases. Critical words were always bisyllabic, and the /s/ and /f/ sound always occurred at the beginning of the second syllable. This was to ensure that the “critical phonemes . . . [were] well-articulated and . . . preceded by relatively strong lexical information” (Kraljic & Samuel, 2005, p. 147). In Kraljic and Samuel’s study, critical sounds occurred at syllable onsets late in words, with most words having three, sometimes four, syllables. Our decision to use bisyllabic words might have reduced the strength of the lexical information preceding the critical sounds (as our results show, this was not an issue), but allowed us to closely match the phonotactic context in critical words for /s/ and /f/ sounds (e.g., *passive-passion*). For the same

Table 1  
Stimuli for S-Label Condition (Experiment 1)

Tongue Twister Context (S-Label)	Non Tongue Twister Context (S-Label)
passion mansion pa? <i>s</i> fivə passion*	holler tamper pa? <i>s</i> fivə holler*
pushing cushion ki? <i>s</i> fɪŋ pushing	kelly bigot ki? <i>s</i> fɪŋ kelly
crucial glacial cla? <i>s</i> fɪk crucial	gecko ruby cla? <i>s</i> fɪk gecko
pension mission po? <i>s</i> fʌm pension	tamer hater po? <i>s</i> fʌm tamer
cashew kosher ca? <i>s</i> fʌlɪt cashew*	layman hating ca? <i>s</i> fʌlɪt hating*
blushing pressure blo? <i>s</i> fɒm blushing*	header leaning blo? <i>s</i> fɒm leaning*
ration washing ran? <i>s</i> fɒm ration*	yapping nodded ran? <i>s</i> fɒm nodded*
bishop gusher go? <i>s</i> fɪp bishop	wacky talent go? <i>s</i> fɪp talent
holler tamper hamper holler*	passion mansion hamper passion*
kelly bigot belly kelly	pushing cushion belly pushing
gecko ruby raking gecko	crucial glacial raking crucial
tamer hater hammer tamer	pension mission hammer pension
layman hating human hating*	cashew kosher human cashew*
header leaning leader leaning*	blushing pressure leader blushing*
yapping nodded napping nodded*	ration washing napping ration*
wacky talent tacky talent	bishop gusher tacky bishop
weary deepen dairy deepen*	weary deepen dairy deepen*
polly gaping goalie gaping*	polly gaping goalie gaping*
carry making marry making	carry making marry making
debit rookie rabbit rookie*	debit rookie rabbit rookie*
hidden berry button berry	hidden berry button berry
bullet happy hamlet happy*	bullet happy hamlet happy*
wacko tamer taco tamer	wacko tamer taco tamer
weary deepen dairy deepen	weary deepen dairy deepen

*Note.* The eight shaded rows in each condition represent the critical stimuli containing a shifted sound (?sʃ). The nonshaded rows in each between-subject condition are the 16 filler phrases. Eight of the filler phrases were identical across all conditions, and contained no fricative sounds. The other filler phrases were balanced between the critical phrases so that participants in the Tongue Twister Context and the Non-Tongue Twister Context of each label condition would hear the exact same recordings. Items marked with an asterisk (\*) represent the subset of items used in Experiment 4 and 5. Our design implies that there are three-times as many unshifted sounds as atypical shifted critical sounds. This differs from previous work and is addressed in Experiment 2.

reason, /s/ and /ʃ/ sounds were always surrounded by either a vowel or nasal sound.

Following previous work, none of the other words contained any other fricative sounds (including /s/ and /ʃ/). Lists for stimuli presentation were created by Latin square design over Label and Context. One pseudorandomized stimulus order (and its reverse) was created in which no more than two critical phrases occurred in a row. This resulted in eight lists (2 Label × 2 Context × 2 Orders = 8 Lists).

We first describe the creation of the shifted pronunciations. We then describe the structure of the critical phrases in the Tongue Twister Context, followed by the structure of the filler phrases in the Tongue Twister Context. Finally, we describe the structure of the critical and filler phrases in the Non-Tongue Twister Context. Phrases were recorded at a natural speech rate, with durations of about 2–2.5 s.

**Creation of shifted pronunciations.** The third word in the phrase was the critical word: the /s/ or /ʃ/ in this word was shifted toward its fricative counterpart (i.e., /ʃ/ or /s/, respectively). To create these atypical productions, the talker (a 25-year-old female, native talker of American English) recorded two versions of each phrase, one containing the normal pronunciation of the third word (e.g., *passive*) and one containing the atypical pronunciation of the third word with the fricative counterpart (e.g., *passive*). The pronunciation containing the fricative counterpart never resulted in a real word, which allowed the participant to use lexical knowledge to disambiguate the identity of the shifted fricative. The /s/ and /ʃ/

of the two recordings were blended using *Fricative Maker Pro* (McMurray, Rhone, & Galle, 2012) to create a continuum with 31 steps for that word (e.g., ranging from *passive* to *passive*). Following Kraljic and Samuel (2005), three native English speakers then independently listened to these words to identify the word that sounded maximally ambiguous. The average of their responses was selected as the shifted /ʃsʃ/ word that was presented to participants. Each shifted pronunciation was then inserted back into the phrases corresponding to the Tongue Twister and Non-Tongue Twister Context, which we describe next.

**Critical phrases in Tongue Twister Context.** We created 8 four-word phrases in each Label condition that contained an atypical pronunciation of either /s/ or /ʃ/ in the third word position (e.g., *passion mansion pa?*s*fivə passion*). Specifically, the phrases were of the structure  $S_1 S_2 ?sʃ S_1$  (or  $f_1 f_2 ?sʃ f_1$ ), where  $S_1$  and  $S_2$  were words that contained the /s/ sound, and ?sʃ was a word that contained the shifted /ʃsʃ/ sound. These tongue twister phrases had a number of structural properties that were intended to make it plausible that they would elicit mispronunciations of /s/ as /ʃ/ (or more /ʃ/-sounding /s/ sounds) in the S-label condition (and vice versa in the f-label condition). For example, the /s/ and /ʃ/ sounds in our experiment all appeared word medially, as speech errors are more likely to affect sounds that share a word position than when they do not (Shattuck-Hufnagel, 1983; Wilshire, 1999).

Additionally, we positioned the atypical /s/ and /ʃ/ sounds in the third word position, preceding a word with a typical pronunciation of the counterpart fricative, because speech errors are likely to

Table 2  
Stimuli for *f*-Label Condition (Experiment 1)

Tongue Twister Context ( <i>f</i> -Label)	Non-Tongue Twister Context ( <i>f</i> -Label)
passive massive pa?sʃion passive*	holler tamper pa?sʃion holler*
kissing missing cu?sʃion kissing	kelly bigot cu?sʃion kelly
classic glassy cru?sʃial classic	gecko ruby cru?sʃial gecko
tossing possum pen?sʃion tossing	tamer hater pen?sʃion tamer
castle missile ca?sʃew castle*	layman hating ca?sʃew hating*
blossom pressing blu?sʃing blossom*	header leaning blu?sʃing leaning*
ransom wussy ra?sʃion ransom*	yapping nodded ra?sʃion nodded*
gossip bicep bi?sʃop gossip	wacky talent bi?sʃop talent
holler tamper hamper holler*	passive massive hamper passive*
kelly bigot belly kelly	kissing missing belly kissing
gecko ruby raking gecko	classic glassy raking classic
tamer hater hammer tamer	tossing possum hammer tossing
layman hating human hating*	castle missile human castle*
header leaning leader leaning*	blossom pressing leader blossom*
yapping nodded napping nodded*	ransom wussy napping ransom*
wacky talent tacky talent	gossip bicep tacky gossip
weary deepen dairy deepen*	weary deepen dairy deepen*
polly gaping goalie gaping*	polly gaping goalie gaping*
carry making marry making	carry making marry making
debit rookie rabbit rookie*	debit rookie rabbit rookie*
hidden berry button berry	hidden berry button berry
bullet happy hamlet happy*	bullet happy hamlet happy*
wacko tamer taco tamer	wacko tamer taco tamer
weary deepen dairy deepen	weary deepen dairy deepen

Note. For details, see caption of Table 1.

anticipate upcoming sounds (Wilshire, 1999). In other words, the third word in our tongue twister phrases were shifted toward the fricative in the first, second, and fourth word of our phrases.

Finally, as much as possible, the first and second word in the phrase shared a common vowel in the second syllable, and either the first or second word in the phrase shared a common onset with the third word. For example, consider the phrase “passive massive pa?sʃion passive.” The first and second words (*passive* and *massive*) share the vowel in the second syllable (in fact, they share the entire syllable), and the first and third word share a common onset (*passive* and *passion*). This stimulus structure was chosen to approximate the type of tongue twisters used in speech error eliciting experiments (e.g., Sevald & Dell, 1994; Shattuck-Hufnagel & Klatt, 1979; Wilshire, 1999). For example, Wilshire (1999) used tongue twisters consisting of four monosyllabic words, where the word-initial phoneme varied in the structure ABBA, and the word-final phoneme varied in the structure ABAB (e.g., *palm neck name pack*).

*Filler phrases in the Tongue Twister Context.* We created 16 four-word filler phrases. In four of these phrases, the first word was repeated in the fourth position (e.g., *holler tamper hamper holler*), and in 12 of these phrases, the second word was repeated in the fourth position (e.g., *weary deepen dairy deepen*). When combined with the eight critical phrases described above this resulted in each participant hearing 12 examples where the first word was repeated in the fourth position, and 12 examples where the second word was repeated in the fourth position. This was done so that participants would not be able to consistently anticipate the fourth word.

Additionally, for each four-word filler phrase, we aimed to select pairs of phonemes to use for the onsets of the first and second syllables of each word that had no (or a very low) incidence

of speech errors with each other, based on the MIT confusion matrix of 1,620 single phoneme errors (Shattuck-Hufnagel & Klatt, 1979). For example, for the filler phrase “holler tamper hamper holler,” both the pairs h/t and l/p are exchanged for each other the fewest number of times in that matrix (0 occurrences).

*Critical and filler phrases in the Non-Tongue Twister Context.* For each Label condition, participants in the Non-Tongue Twister Context heard exactly the same recordings of words as those in the Tongue Twister Context. To achieve this, for each of the critical phrases in the Tongue Twister Context, we spliced the third word (containing the shifted /ʃsʃ/) into one of the filler phrases. For example, in the S-Label/Tongue Twister condition, one of the critical phrases is “passion mansion pa?sʃive passion” and one of the filler phrases is “holler tamper hamper holler.” In the S-Label/Non-Tongue Twister condition, the critical phrase becomes “holler tamper pa?sʃive holler” and the filler phrase becomes “passion mansion hamper passion” (see Tables 1 and 2 for full list of stimuli). Thus, in the Tongue Twister Context, the word containing a shifted /ʃsʃ/ occurs in a phrase that was created to make speech errors seem plausible, while in the Non-Tongue Twister Context, it does not.

*Test block: Categorization task.* Following previous work, the same talker who recorded the exposure stimuli was recorded saying the nonce words /asi/ and /afi/. These nonce words were blended together using FricativeMakerPro (McMurray et al., 2012) to create a continuum of 31 steps ranging from /asi/ to /afi/. We selected seven of these steps to serve as test steps based on initial informal piloting: five steps were centered close to the point of maximal ambiguity, and two steps represented category endpoints. This procedure closely follows previous work, though the specific numbers of test tokens and their placement along the continuum varies somewhat across works (e.g., Kraljic & Samuel, 2006; Liu

& Jaeger, 2018; Norris et al., 2003; Vroomen, van Linden, De Gelder, & Bertelson, 2007). The results reported in Figure 3 below confirm that the seven test steps span the /s-/ʃ/ continuum, as intended.

**Procedure.** The experiment began with instructions, one practice trial, the exposure block, the test block, and the postexperimental survey. This general structure was identical to that used in many previous perceptual recalibration experiments. Previous work has often used lexical decision tasks during exposure (e.g., Kraljic & Samuel, 2005; Liu & Jaeger, 2018; Norris et al., 2003; Zhang & Samuel, 2014), though perceptual recalibration has also been found for a broad variety of different tasks during exposure (e.g., passive listening with catch trials, e.g., Bertelson, Vroomen, & De Gelder, 2003; Vroomen et al., 2007; ABX discrimination, Clarke-Davidson, Luce, & Sawusch, 2008; categorization, Clarys, Tanenhaus, Aslin, & Jacobs, 2008; Kleinschmidt et al., 2015; for further review and comparison of various paradigms, see Drouin & Theodore, 2018). Here we use a transcription task during exposure, a paradigm often used in related work on accent adaptation (e.g., Baese-Berk, Bradlow, & Wright, 2013; Bradlow & Bent, 2008; Tzeng, Alexander, Sidaras, & Nygaard, 2016; Xie, Liu, & Jaeger, 2019).

During the practice trial, participants heard a male, native American-English accented talker saying the words “grumpy kitten table pretty.” This talker was clearly different from the exposure and test talker, who was female. After the phrase, participants were asked to transcribe the words that they heard, separated by spaces. The trial was repeated until participants correctly transcribed the words. The purpose of the practice trial was to familiarize participants with the task and to allow them to adjust the volume to a comfortable listening level.

The exposure block trials followed the exact same format as the practice trial, except that participants did not receive feedback on the accuracy of their transcriptions and each trial was only played once. Participants heard 24 trials, separated by an intertrial interval of 1,000 ms. With a total of 96 words (24 four-word trials), the amount of exposure was similar to our previous web-based studies in which we found perceptual recalibration for /s-/ʃ/ (e.g., 60–160 trials across the three experiments reported in Liu & Jaeger, 2018).

One difference of the current study to the more common lexical decision paradigm is the relative proportion of unshifted and atypical pronunciations. In the present study, because of the necessary repetition of sounds in the Tongue Twister Context, participants heard three times as many nonshifted pronunciations (of /s/ or /ʃ/) as shifted pronunciations, whereas previous studies have exposed participants to equal number of unshifted and atypical pronunciations. Most accounts of perceptual recalibration would predict that the degree of boundary shift primarily depends on the number of atypical pronunciations (Experiment 2 assesses and confirms this assumption). The number of atypical pronunciation and their relative proportion out of all trials in the present study (eight critical atypical items out of 96, i.e., 8.3%) was similar compared with previous web-based studies in which we found perceptual recalibration for /s-/ʃ/ (e.g., 6–16 atypical items at a rate of 10% of all items, in Liu & Jaeger, 2018).

During the test block participants, categorized seven steps on the /asi-/afi/ continuum as either /asi/ or /afi/, five times each. The steps were played in five cycles (trial bins), each containing a random ordering of the seven steps. Participants indicated their

responses using the ‘X’ and ‘M’ keys on their keyboard. Key bindings were counterbalanced across participants. This test procedure is identical to that of our previous web-based studies in which we found perceptual recalibration for /s-/ʃ/, except that we halved the number of trials bins from 10 to 5. We did so because our previous work found that the perceptual recalibration effect is largest at the beginning of the test block and then steadily decreases (Liu & Jaeger, 2018; confirmed below).

Finally, participants answered a questionnaire that asked about their audio equipment, language background, technical difficulties, and attention during the experiment.

**Scoring transcription accuracy during exposure.** An undergraduate research assistant compiled a list of common misspellings for each word (e.g., spelling *polly* as *pollie* or *poly*). Transcription accuracy was automatically scored for matches to the expected transcriptions; any word that was also on the list of common misspellings was labeled as correct. We counted a word’s transcription as correct regardless of whether the four words had been transcribed in the correct order. The same scoring approach was used for all other experiments reported below. If word order mistakes were counted, transcription accuracies would decrease by about 7.8% across all experiments (range = 5.8–12.1%). None of the results reported in this paper change if order mistakes are counted (for full information, see online supplemental data information).

## Results

We first summarize our analyses of transcription accuracy during the exposure block. We then turn to the critical results from the test block, comparing the Label and Context conditions. We use mixed logistic regression (Breslow & Clayton, 1993; Jaeger, 2008) to analyze responses during both the exposure and the test phase, as both involve binary dependent variables.

**Exposure block: Transcription accuracy.** Following previous work, we analyzed two aspects of transcription accuracy during exposure. We first examined the overall accuracy to assess whether participants were listening to the stimuli. Then, we assessed whether participants transcribed the critical shifted words correctly. A failure to do so *might* suggest that participants did not recognize the words, which in turn might reduce the magnitude of perceptual recalibration. Figure 2 summarizes the overall transcription accuracy for all experiments.

For Experiment 1, the overall transcription accuracy averaged over by-participant means was 88.4% ( $SD = 7.0\%$ ). Table 3 shows accuracies by between-participants conditions. Given the challenging task of transcribing 24 sequences of four semantically unrelated words spoken at a rate of about two words per second, we take this to be adequate performance, indicating that participants were paying attention during the exposure block. To assess whether transcription accuracy differed between conditions, we conducted a mixed logit regression predicting trial-level accuracy (1 = correct, 0 = incorrect) from Label (always sum-coded: f-Label = 1 vs. S-Label = -1), Context (sum-coded: Non-Tongue Twister = 1 vs. Tongue Twister = -1), and their interaction. The analysis included by-participant intercepts and by-item random intercepts and slopes for Context, Label, and their interaction. An item was defined as the  $n$ th row of Tables 1 and 2. For



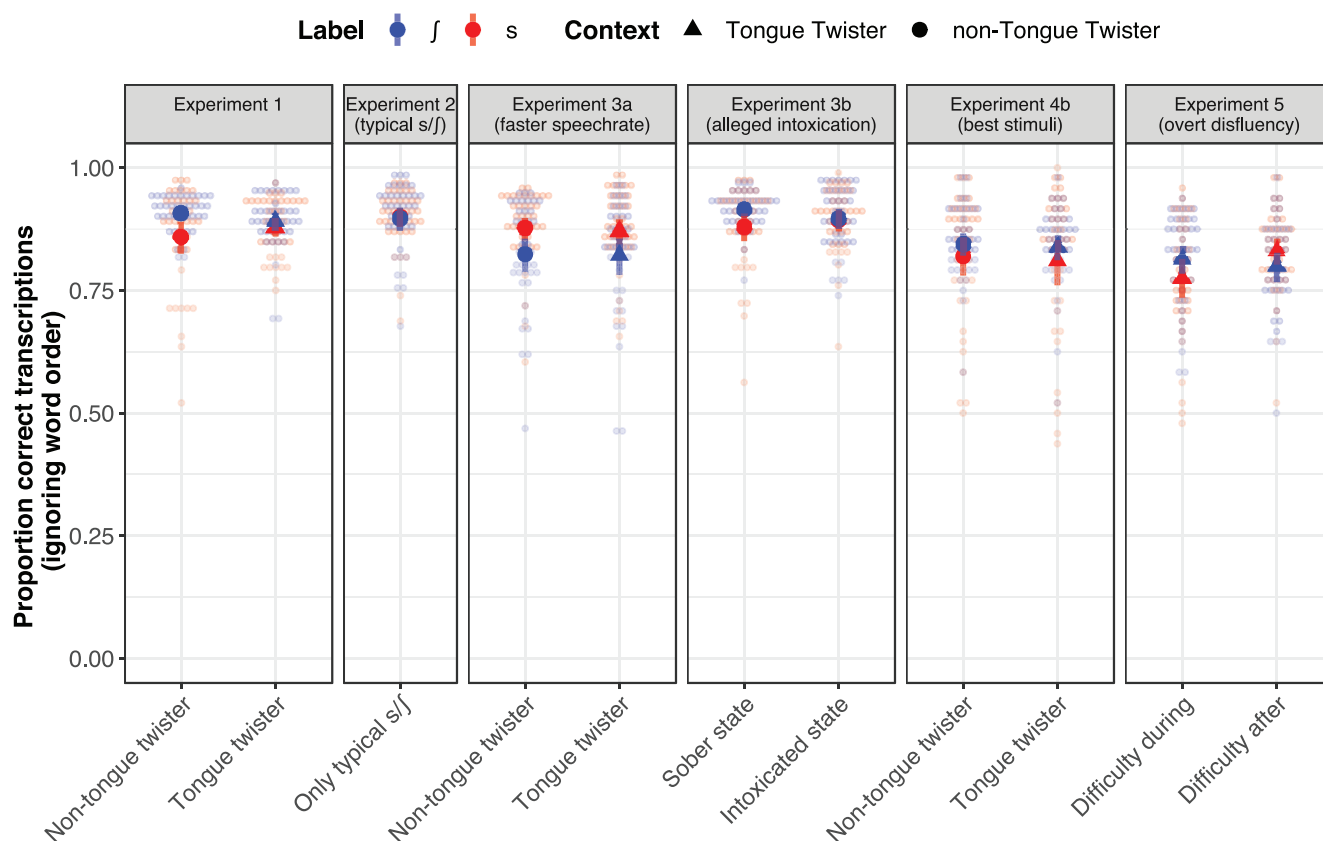


Figure 2. Transcription accuracy during exposure for all experiments and between-participants conditions. Transparent points show by-participant averages. Solid point ranges show the mean and bootstrapped 95% confidence interval over those by-participant averages. See the online article for the color version of this figure.

example, the first rows of Tables 1 and 2 together constitute one item.

Participants in the *f*-Label condition transcribed significantly more words correctly than those in the *s*-Label condition ( $\hat{\beta} = 0.27$ ,  $z = 3.11$ ,  $p < .002$ ). This effect was small (see Table 3), and does not hold across all experiments (see Figure 2). To foreshadow the results of Experiments 2–5, Experiment 3b exhibited the same effect as Experiment 1 ( $\hat{\beta} = 0.18$ ,  $z = 2.21$ ,  $p < .03$ ), but Experiment 3a exhibited a similarly small effect in the opposite direction—*lower* accuracy in the *f*-Label condition ( $\hat{\beta} = -0.20$ ,  $z = -2.61$ ,  $p < .01$ ). No main effects of Label condition were observed in Experiments 2, 4b, and 5. Given our interest in tongue twister contexts, neither the effects of Context ( $p > .92$ ), nor its interaction between with Label was significant ( $p > .48$ ). It is unlikely that any effect of Context (or lack thereof) on the categorization boundary during the test phase could be confounded by overall task engagement.

Next, we analyzed the proportion of correctly transcribed critical shifted words with the exact same analysis approach. The mean transcription accuracy of shifted words in Experiment 1 was 84.8% ( $SD = 16.3$ ). Table 4 shows accuracies by between-participants conditions. The mixed logit regression found that participants in the *f*-Label condition transcribed significantly more shifted words correctly than those in the *s*-Label condition ( $\hat{\beta} = 0.56$ ,  $z = 4.28$ ,

$p < .0001$ ). This effect was larger than for overall accuracy, possibly driving the effects on overall accuracy. Neither the effects of Context ( $p > .85$ ), nor its interaction between with Label was significant ( $p > .84$ ). The same holds for all experiments reported below: none of our Context manipulations had a significant main effect on the overall accuracy, or the accuracy with which shifted tokens were transcribed; similarly, Context never interacted significantly with the Label condition (though there were marginally significant interactions in Experiments 3b and 5).

In short, there is no reason to expect that differences in task engagement during exposure, or differences in participants' ability to recognize and process the shifted words would confound the analyses of the test data reported below. Still, to address our own concerns and those of reviewers', additional control analyses are reported in the online supplemental data, available at <https://osf.io/ungba/>. Specifically, we repeated all analyses of category boundary shifts during the test block (for all experiments) while also including the participant's transcription accuracy during exposure as a predictor, as well as all interactions of that predictor with all other variables in the analysis. All of these analyses confirmed the results we report below: while higher accuracy during exposure predicted larger perceptual recalibration effects during test in some of the experiments, this effect never changed



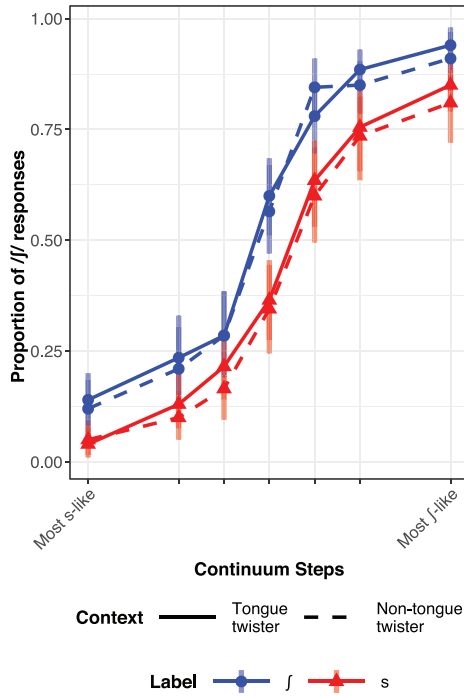


Figure 3. Proportion of /f/ responses as a function of Continuum Step (Experiment 1). Participants in the f-Label condition (blue triangle) shift toward /s/ and participants in the S-Label condition (red circle) shift toward /f/ for both Context conditions. Error bars show 95% confidence intervals obtained via nonparametric bootstrap over the by-participant means. Note that our analysis (unlike this figure) follows previous work and collapses across continuum steps. See the online article for the color version of this figure.

the significance of Label or its interaction with Context. This was the case regardless of the specific accuracy measure used.

**Test block: Changes in the categorization boundary.** We present two planned analyses. Both analyses are trial-level analyses over all data from the test block. The first analysis follows standard practice, and analyzes the average proportion of /f/ responses ignoring continuum steps and trial order. This resembles the analyses of variance presented in the majority of studies on perceptual recalibration. The second analysis assesses the perceptual recalibration effect at the *beginning* of the test block. The reason for this second (planned) analysis is found in our previous work: in Liu and Jaeger (2018) we found that perceptual recalibration effect continuously reduced during the test block, perhaps because of the uniform distribution of stimuli across the /asi-/afsi/

Table 3  
Transcription Accuracy by Label and Context Condition With Standard Deviations in Parentheses (Experiment 1)

	Percent of words correctly transcribed	
	f-Label	S-Label
Nontongue twister	90.7% (3.6%)	85.9% (10.7%)
Tongue twister	89.1% (6%)	87.8% (5.9%)

Table 4  
Transcription Accuracy for Only the Eight Critical Shifted Words (Experiment 1)

	Percent of words correctly transcribed	
	f-Label	S-Label
Nontongue twister	90.6% (10.9%)	83.1% (22.6%)
Tongue twister	83.4% (13.7%)	82.2% (14.9%)

continuum. This means that the standard analysis—assessing average perceptual recalibration across the entire test block—can substantially underestimate the true perceptual recalibration (as we confirm this below). Such undoing of perceptual recalibration effects during testing is expected if perceptual recalibration reflects distributional learning (as argued in, e.g., Kleinschmidt & Jaeger, 2015; Lancia & Winter, 2013).

Our second analysis directly addresses this possibility by capturing changes in perceptual recalibration during test, and providing a measure of perceptual recalibration at the beginning of the test block. As we show below, this increases our ability to detect effects on perceptual recalibration (such as the hypothesized blocking of perceptual recalibration). Following our previous work, all subsequent analyses are based on this alternative approach.

For the Non-Tongue Twister Context we predict the same type of perceptual recalibration as in previous studies with different exposure tasks (Kraljic & Samuel, 2005, 2011; Kraljic et al., 2008; Liu & Jaeger, 2018; Norris et al., 2003): participants in the f-Label should shift their category boundary toward /s/ and, thus, categorize more sounds as /f/, and participants in the S-Label shift their category boundary toward /f/ and, thus, should categorize more sounds as /s/.

In the Tongue Twister Context, participants heard the same shifted pronunciations as in the Non-Tongue Twister Context, but embedded in a tongue twister. If the tongue twister context provided participants with a plausible causal explanation for the atypical pronunciations, then participants may attribute these atypical pronunciations to an incidental cause, leading them to adapt less or not at all (as observed for visually provided cause in Kraljic & Samuel, 2011; Kraljic et al., 2008; Liu & Jaeger, 2018).

**Average perceptual recalibration across the test block.** Figure 3 shows the categorization curve for all four conditions of Experiment 1 (averaged across all trial bins). We conducted mixed logit regression, where we predicted /f/ responses (1 = /f/ response, 0 = /s/ response) by Label (sum-coded: f-Label = 1 vs. S-Label = -1) and Context (sum-coded: Non-Tongue Twister = 1 vs. Tongue Twister = -1), and their interaction. The analysis included by-participant random intercepts, which constitutes the maximal random effect structure for our design.

This revealed that overall more /f/ responses were observed in the f-Label condition than in the S-Label condition ( $\beta = 0.30, z = 5.0, p < .001$ ; Figure 3). This is consistent with perceptual recalibration, and a shift in the categorization boundary based on Label. The output from the model is shown in Table 5. Critically, there was no significant difference of Context ( $p = .41$ ) nor was there a significant interaction between Label and Context ( $p = .69$ ). This suggests that participants who heard the atypical pro-

Table 5  
Mixed Logit Regression Predicting Proportion of /f/ Responses  
From Label, Condition, and Their Interaction (Experiment 1)

Predictors	Parameter estimates		Significance test	
	Coef ( $\hat{\beta}$ )	SE	z	p
(Intercept)	-.10	.06	-1.62	.10
Label (f vs. S)	.30	.06	5.0	<b>&lt;.001</b>
Context (NonTT vs. TT)	-.05	.06	-.8	.41
Label:Context	.02	.06	.40	.69

Note. Coding: Label (sum coded: f-Label = 1 vs. S-Label = -1) and condition (Non-Tongue Twister Context = 1 vs. Tongue Twister Context = -1). Here and in all other result tables, rows that are critical to our analysis are highlighted in grey. Significant effects are shown in bold, marginal effects in italics.

nunciations in the Tongue Twister Context adapted just as strongly as those who had heard these pronunciations in the Non-Tongue Twister Context, contrary to what we had originally predicted.

**Measuring perceptual recalibration at the beginning of test block.** Replicating Liu and Jaeger (2018), we find that participant responses move toward a 50/50 (empirical logit of 0) asi/-af/i/

baseline over the course of the test block (see Figure 4). Following Liu and Jaeger (2018), we conducted an additional analysis to assess the perceptual recalibration effect at the very beginning of the test block. We used mixed logit regression to predict /f/ responses from Label (sum-coded: f-Label = 1 vs. S-Label = -1), Context (sum-coded: Non-Tongue Twister = 1 vs. Tongue Twister = -1), Trial Bin (coded continuously with the first trial bin as 0), and their interactions (see Table 6). The estimated effect of Label represents the estimate of the recalibration effect across both Context conditions during the first trial bin of the test block. This and all subsequent analyses of this type included by-participant random intercepts (models with by-participant random slopes for Trial Bin did not converge or led to singular fits, except for Experiment 3a).

We again found that participants in the f-Label condition provided more /f/ responses than those in the S-Label condition ( $\hat{\beta} = 0.56$ ,  $z = 7.59$ ,  $p < .001$ ). That is, the true perceptual recalibration effect at the beginning of the test block (1.12 log-odds =  $\hat{\beta} \times 2$  since we used -1 vs. 1 sum-coding) is almost twice as large as the estimate one obtains from averaging across the entire test block (0.6 log-odds). This validates the need for the advanced analysis, which we continue to use throughout the remainder of the article.

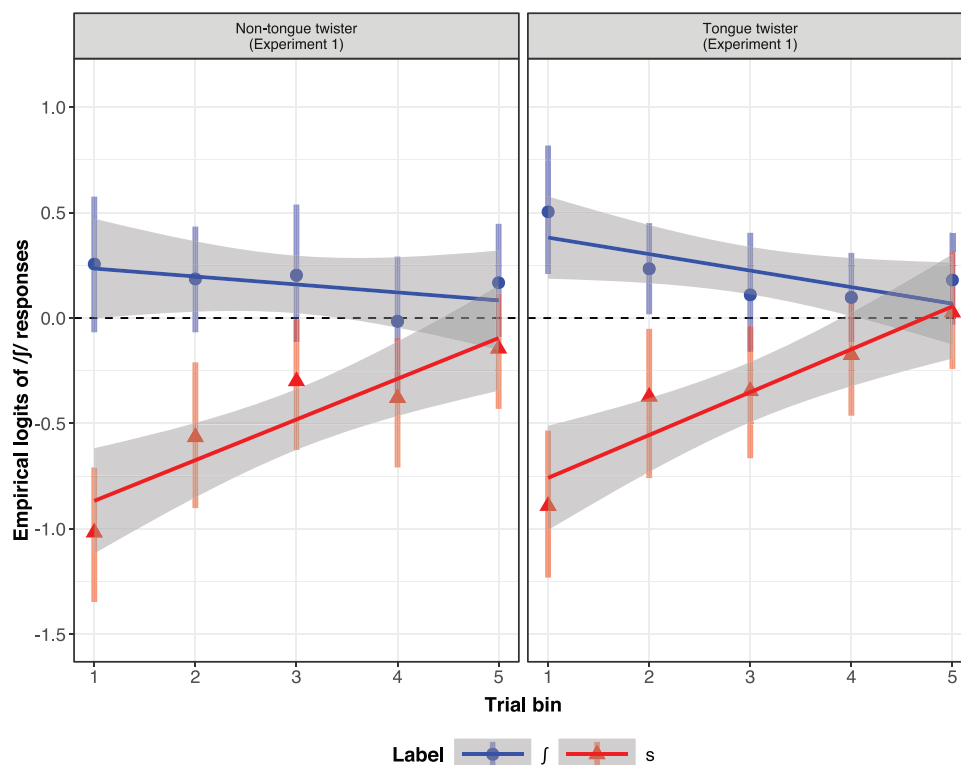


Figure 4. Proportion of /f/ responses as a function of exposure condition and Trial Bin (Experiment 1). Proportions of /f/ responses were empirical logit transformed to facilitate comparison with the model's prediction. Point ranges show empirical means of by-participant averages, and 95% bootstrapped confidence intervals over those by-participant averages. Solid lines show predictions of the model used to obtain corrected estimates of the category boundary shift at the beginning of the test block. Over the course of testing, categorization responses in all exposure conditions move toward 0 empirical logits (i.e., 50/50 /s/ and /f/ responses, dashed line). See the online article for the color version of this figure.

Table 6  
Mixed Logit Regression Predicting Proportion of /f/ Responses  
From Label, Condition, and Their Interaction (Experiment 1)

Predictors	Parameter estimates		Significance test	
	Coef ( $\hat{\beta}$ )	SE	z	p
(Intercept)	-.25	.07	-3.37	<b>&lt;.001</b>
Label (f vs. S)	.56	.07	7.59	<b>&lt;.001</b>
Context (NonTT vs. TT)	-.07	.07	-1.00	.32
Trial Bin (first bin = 0)	.07	.02	3.63	<b>&lt;.001</b>
Label:Context	.01	.07	.19	.85
Label:TrialBin	-.13	.02	-6.27	<b>&lt;.001</b>
Context:TrialBin	.01	.02	.56	.57
Label:Context:TrialBin	.01	.02	.27	.79

Note. Coding: Label (sum coded: f-Label = 1 vs. S-Label = -1), condition (Non-Tongue Twister Context = 1 vs. Tongue Twister Context = -1), trial bin (first bin = 0). Rows that are critical to our analysis are highlighted in grey. Significant effects are shown in bold.

The total number of /f/ responses tended to increase over trial bins ( $\hat{\beta} = 0.07$ ,  $z = 3.65$ ,  $p < .001$ ), and that this differed between Label conditions, in a way consistent with convergence toward 50/50: participants in the f-Label condition tended to provide fewer /f/ responses in later trial bins, compared with those in the S-Label condition ( $\hat{\beta} = -0.13$ ,  $z = -6.27$ ,  $p < .001$ ). This behavior is clearly visible in Figure 4.<sup>2</sup>

Critically, the interaction between Label and Context was again nonsignificant ( $p = .85$ ), suggesting that even in the first trial bin there was no evidence that the effect of perceptual recalibration differed depending on whether the shifted pronunciations were embedded in a tongue twister context or not (see Figure 4).

## Discussion

In Experiment 1, we find that exposure to atypical pronunciations of /s/ or /f/ from one talker leads participants to change how they categorize sounds on the /s/-/f/ continuum. Specifically, participants who are exposed to words containing atypical sounds labeled as /f/ then categorize more sounds as /f/, leading to a shift in their categorization curve toward /s/. This replicates the results of previous perceptual recalibration studies, but in a novel multi-word phrase transcription paradigm. The size of the perceptual recalibration effect at the beginning of the test block was comparable with previous work. Specifically, Liu and Jaeger (2018) found a perceptual recalibration effect (the difference between the two Label conditions) of 1.65 log-odds for 10 critical tokens (Experiment 1) and a perceptual recalibration effect of a little under 1.0 log-odds for six critical tokens (Experiment 2).<sup>3</sup> The present result of 1.12 log-odds for eight critical tokens falls within the expected range. This provides initial validation of the present paradigm, as most theories of perceptual recalibration would predict the effect to increase with the number of critical tokens.

Our paradigm differs from the standard perceptual recalibration paradigm in that participants heard three times as many typical pronunciations as they heard shifted pronunciations (e.g., participants in the f-Label condition heard 24 typical /s/, and eight shifted /f/). This contrasts with previous experiments, where participants typically heard equal numbers of typical and shifted pronunciations.

One potential concern is that the increased number of typical pronunciations that participants heard might have affected how participants categorized sounds, and that this overrides any potential effect of causal attribution—and, thus, the hypothesized effect of tongue twisters on the adaptation to atypical pronunciations. For example, work on selective adaptation has found that repeated presentations of typical instances of one phoneme leads listeners to categorize fewer sounds as that same phoneme. This effect has variously been attributed to the fatigue of “linguistic feature detectors” or other phonetic assignment processes (Eimas & Corbit, 1973; Samuel, 1986), or a shrinking of the variance for the listener’s underlying distribution for that phonetic category or a change in the prior probability for a category (Kleinschmidt & Jaeger, 2016).

Two considerations ameliorate this concern. First, there are striking differences between the present paradigm and selective adaptation studies. For example, selective adaptation paradigms tend to repeat the typical stimulus many dozens of times (e.g., Samuel, 1989; Vroomen et al., 2007). Second, we observe effect sizes that match what is expected under previous perceptual recalibration experiments. This would be rather unexpected, if the differences in paradigms had a large effect on our results.

Still, it is theoretically possible that the shift in categorization boundary in Experiment 1 is driven by the repeated typical sounds, rather than the shifted sounds. In that case, no effect of Context is expected (both context conditions contained equally many unshifted typical sounds, and Tongue Twister contexts are not expected to block the effect of exposure to typical pronunciations). We decided to conduct Experiment 2 to directly address the possibility that the shift in the categorization boundary in Experiment 1 was driven entirely by the repetition of unshifted phonemes. As we detail below, Experiment 2 also serves as a baseline to both Experiment 1 and the subsequent experiments.

## Experiment 2

Experiment 2 tests whether the shift in the categorization boundary for the /s/ and /f/ phonemes that we observed in Experiment 1 could be driven solely by the higher number of typical unshifted, compared with atypical shifted, tokens. To this end, we presented participants with only typical, unshifted instances of /s/ and /f/ (interspersed with the same filler trials as in Experiment 1). These typical sounds were presented in the same Non-Tongue Twister Contexts used in Experiment 1, so that the only difference to Experiment 1 was that the atypical pronunciations were replaced with unshifted productions. In Experiment 2, participants in the f-Label group heard three times as many unshifted /s/ as unshifted

<sup>2</sup> We note an important—at first blush potentially counter-intuitive—methodological consequence that also applies to the study of other adaptation and learning phenomena: longer test blocks, intended to collect more test data to increase statistical power, can actually result in less power to detect learning effects if the test block is structured in a way that leads to undoing of the learning effect, and analyses do not take into account that learning might continue throughout the test block (for discussion, see Jaeger, 2010, p. 53; Jaeger, Burchill, & Bushong, 2019).

<sup>3</sup> The analyses in Liu and Jaeger (2018) used (.5 vs. -.5) sum-coding, whereas the present analyses use (1 vs. -1) sum-coding. Whenever we compare effect sizes below, we adjust for this difference (that does not affect significance testing).

/f/, and participants in the S-Label group heard three times as many unshifted /f/ as unshifted /s/.

If Experiment 2 finds the same shift in the categorization boundary as Experiment 1, this would constitute strong evidence that the effect observed in Experiment 1 is, in fact, not because of perceptual recalibration (because Experiment 2 does not contain any atypical pronunciations of /s/ or /f/). However, if we fail to find a difference between the two Label conditions in Experiment 2 or if the difference is weaker than it is in Experiment 1, this would suggest that the effect from Experiment 1 is at least in part because of perceptual recalibration. This in turn would raise the question why the perceptual recalibration effect in Experiment 1 is not blocked by the presence of an incidental cause.

## Method

**Participants.** We recruited 86 participants on Amazon Mechanical Turk (MTurk) for a target of 40 participants in each of the two Label conditions after exclusions. Two participants were excluded for not correctly identifying the speaker as female, one participant for not wearing headphones, and three for inverted categorization functions. As in Experiment 1, participants were paid \$1 for this experiment, which took roughly 10 min (\$6/hr).

**Materials.** The stimuli used during the exposure block were identical to the stimuli from the Non-Tongue Twister Context of Experiment 1, except that we substituted the critical atypical pronunciation with the typical pronunciation of the same word. These typical pronunciations were the endpoints of the continuum used to create the atypical pronunciations that contained shifted /ʔsʃ/. For example, in the Non-Tongue Twister Context condition of Experiment 1, participants would hear “holler tamper paʔsʃive holler,” but in the current condition, participants would hear “holler tamper passive holler.” All other stimuli were identical. We refer to the current conditions as “Unshifted” conditions, and the Non-Tongue Twister Context conditions from Experiment 1 as “Shifted” conditions.

**Procedure.** The procedure was identical to Experiment 1.

## Results

We first analyze the results from the test block to assess whether participants in the S-Label condition differed in how they categorized sounds compared with participants in the f-Label condition. We next compare the difference between Label conditions in the current experiment (Unshifted condition) with the difference in Label conditions (perceptual recalibration effect) identified in the Non-Tongue Twister Context of Experiment 1 (Shifted condition). Taken together our results suggest that the effect in Experiment 1 is unlikely to be due solely to the higher number of typical pronunciations and, thus, is likely to reflect perceptual recalibration.

The transcription accuracies for Experiment 2 and all subsequent experiments are summarized in Figure 2. With an overall accuracy of 89.8% ( $SD = 6.5\%$ ), transcriptions in Experiment 2 were similar to Experiment 1 (88.4%). For details, see online supplemental data.

**Test block: Changes in the categorization boundary.** First, we assessed whether there was a difference in categorization between the two Label conditions. To do this, we conducted mixed

logit regression, where we predicted categorization by Label condition (sum-coded: f-Label = 1 vs. S-Label = -1), Trial Bin (First bin = 0), and their interaction. This analysis is presented in Table 7.

We did not identify a significant effect of Label at the first Trial Bin ( $\hat{\beta} = 0.17, z = 1.49, p = .14$ ). Numerically, the effect trended in the same direction as the significant effect in Experiment 1, though it was much smaller (0.34 log-odds in Experiment 2, compared with 1.12 in log-odds Experiment 1). To more directly test whether the effect in Experiment 1 could have been caused simply by the higher proportion of typical pronunciations, we compared the unshifted condition (Experiment 2) to the Shifted condition (Non-Tongue Twister condition of Experiment 1). These two conditions are identical with the exception that the critical words contained unshifted /s/ or /f/ instead of the shifted /ʔsʃ/. We conducted a mixed logit regression predicting categorization by Label condition (sum-coded: f-Label = 1 vs. S-Label = -1), Shift (sum-coded: Shifted = 1 vs. Unshifted = -1), Trial Bin (First bin = 0), and their interactions. The results are presented in Table 8 and visualized in Figure 5.

Participants in the f-Label condition tended to categorize more sounds as /f/ ( $\hat{\beta} = 0.37, z = 4.67, p < .001$ ). Critically, there was a significant interaction between Label and Shift ( $\hat{\beta} = 0.21, z = 2.6, p < .01$ ), and participants who heard shifted critical words categorized significantly fewer sounds as /f/ than those who heard Unshifted critical words ( $\hat{\beta} = -0.17, z = -2.13, p < .05$ ). Simple effects analysis confirmed that the Label condition had an effect in Experiment 1 ( $\hat{\beta} = 0.58, z = 5.10, p < .0001$ ) but not Experiment 2 ( $\hat{\beta} = 0.17, z = 1.48, p > .14$ ).

## Discussion

The results of Experiment 2 suggest that it is unlikely that the effects of the Label condition in Experiment 1 originate solely in the larger proportion of unshifted, compared with shifted, pronunciations. This result is not unexpected given differences between the current experiment and paradigms used to study selective adaptation. Experiments on selective adaptation tend to repeat the typical pronunciation many dozens of times (e.g., Bowers, Kazanina, & Andermane, 2016; Samuel, 1989, 1997; Vroomen et al., 2007). By contrast, in the current experiment, the repeated typical sounds totaled only 24 tokens. It would have been surprising to see large selective adaptation effects as driving the effects in Experiment 1. Experiment 2 confirmed this.

Table 7  
Mixed Logit Regression Predicting Proportion of /f/ Responses From Label, Condition, and Their Interaction (Experiment 2)

Predictors	Parameter estimates		Significance test	
	Coef ( $\hat{\beta}$ )	SE	z	p
(Intercept)	.01	.11	.11	.91
Label (f vs. S)	.17	.11	1.49	.14
TrialBin (first bin = 0)	.02	.03	.55	.59
Label:TrialBin	-.05	.03	-1.7	=.09

Note. Coding: Label (sum coded: f-Label = 1 vs. S-Label = -1), trial bin (first bin = 0). Rows that are critical to our analysis are highlighted in grey.



Table 8  
Mixed Logit Regression Predicting Proportion of /f/ Responses From Label, Condition, and Their Interaction (Experiments 1 and 2)

Predictors	Parameter estimates		Significance test	
	Coef ( $\hat{\beta}$ )	SE	z	p
(Intercept)	-.16	.08	-1.98	= .05
Label (f vs. s)	.37	.08	4.66	<.001
Shift (shifted vs. unshifted)	-.17	.08	-2.13	<.05
TrialBin (first bin = 0)	.05	.02	2.5	<.01
Label:Shift	.21	.08	2.59	<.01
Label:TrialBin	-.09	.02	-4.21	<.001
Shift:TrialBin	.04	.02	1.73	= .08
Label:Shift:TrialBin	-.04	.02	-1.84	= .07

Note. Coding: Label (sum coded: f-Label = 1 vs. s-Label = -1), shift condition (sum-coded: shifted critical pronunciation (Experiment 1) = 1 vs. unshifted pronunciations (Experiment 2) = -1), trial bin (first bin = 0). Rows that are critical to our analysis are highlighted in grey. Significant effects are shown in bold, marginal effects in italics.

Experiment 2 also serves as a baseline for Experiment 1 and all subsequent experiments we report: in Experiment 2, participants were exposed to only typical sounds and never heard shifted sounds. A comparison of the left panel in Figure 5 (Experiment 2) to the right panel (the Non-Tongue Twister condition in Experiment 1) suggest that the perceptual recalibration effect is driven by

only the S-Label condition. This was confirmed by a simple effects analysis comparing the two conditions: participants who were exposed to shifted S-Label words in Experiment 1 identified significantly fewer sounds as /f/ than those who had been exposed to unshifted S-Label words in Experiment 2 ( $\hat{\beta} = -0.38$ ,  $z = -3.23$ ,  $p < .01$ ); in contrast, there was no difference between Experiments 1 and 2 for participants in the f-Label condition ( $p = .82$ ). The same asymmetry was found when comparing Experiment 1 to an alternative baseline experiment (reported as Experiment 2b in the online supplemental data). In the alternative baseline experiment participants were exposed only to filler phrases—that is, the complete absence of any /s/ or /f/ during exposure—and then measured category shifts during the same test phase as in Experiments 1 and 2. The categorization boundary observed in that experiment was identical to that observed in Experiment 2.

This asymmetry differs from previous experiments in which we identified perceptual recalibration away from the baseline for both /s/ and /f/ (Liu & Jaeger, 2018). Similar asymmetries have, however, been observed in other work (e.g., Drouin, Theodore, & Myers, 2016; Zhang & Samuel, 2014). Indeed, which of two sound categories elicits perceptual recalibration can differ between experiments (for review, see Samuel, 2016, p. 111), possibly because of stimulus-specific properties and, in particular, the placement of the test continuum relative to the acoustic properties of exposure tokens (for evidence and discussion, see Drouin et al., 2016).

Finally, Experiment 2 further ameliorates concerns that Experiment 1 may suffer from lack of power to detect effects of tongue

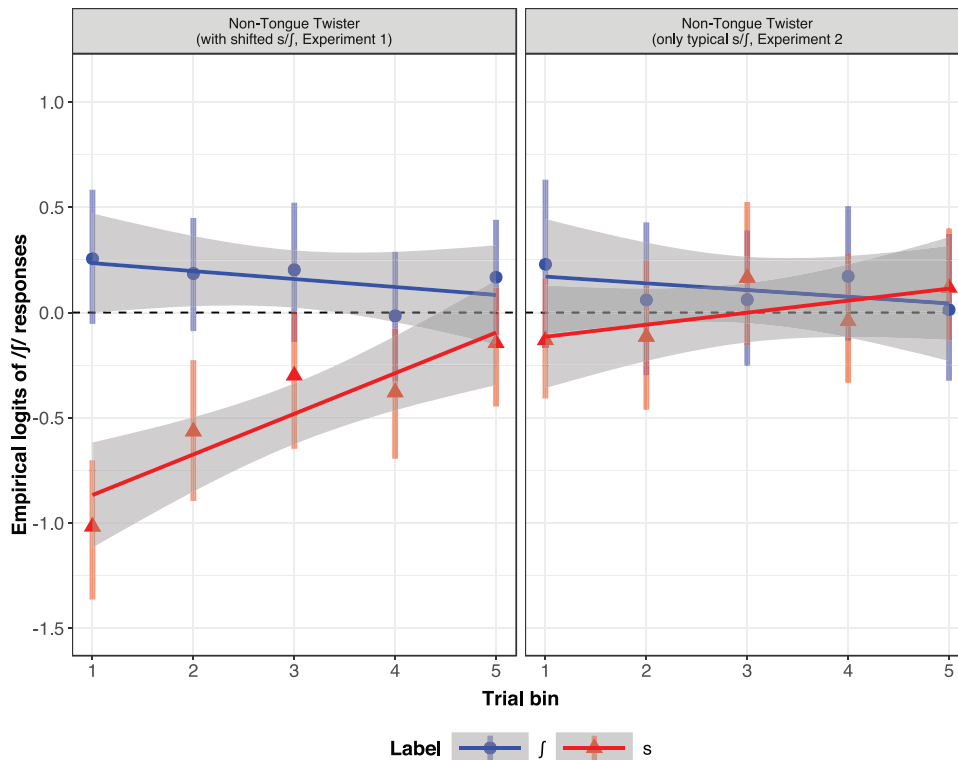


Figure 5. Empirical logits of /f/ responses as a function of Trial Bin (Experiments 1 and 2). For further information, see caption of Figure 4. To facilitate comparison across experiments, the range of the y-axes is held constant here and in all other result plots. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

twister contexts. Our power simulations (see Appendix) found more than 95% power to detect the effect of Label and more than 80% power to detect blocking of those effects. Experiment 2 shows that we can indeed detect the absence or blocking of a perceptual recalibration effect, compared with Experiment 1.

### Experiments 3a and 3b

One possibility for why the tongue twister context did not block adaptation in Experiment 1 is that the tongue twister context we provided was not a sufficiently plausible cause of the atypical pronunciations for participants. If participants did not view the tongue twister context to cause production difficulties, in the way a real tongue twister would, then it is not surprising that we do not find blocking of adaptation when shifted pronunciations are presented in this context.

In the current experiment, we address this possibility in two ways. In Experiment 3a, we increase the plausibility that our tongue twisters would be viewed as tongue twisters, as intended. Production experiments have found an increased incidence of errors when speech rate is increased (MacKay, 1982). To increase the plausibility that our tongue twisters would be viewed as likely to have caused the atypical pronunciations, we increase the speech rate of our stimuli. Additionally, we provide participants in the Tongue Twister Context with explicit information stating that they will hear tongue twisters that may have been difficult for the talker to produce. Explicit instructions of this type have sometimes been found to facilitate attribution to alternative causes, such as intended here (e.g., Arnold et al., 2007, discussed below). In Experiment 3b, we provide participants with an alternative (nontongue twister) cause for the atypical pronunciations. We inform participants that the talker is intoxicated. These experiments taken together allow us to assess whether inferences about causes during speech perception may be influenced by explicit instructions.

### Experiment 3a

#### Method

**Participants.** We recruited 177 participants on MTurk to achieve a target of 40 participants in each of four conditions (S-Label/*f*-Label × Tongue Twister/Non-Tongue Twister). Seven participants were excluded for not correctly identifying the speaker as female and 10 participants for not wearing headphones (9.6% exclusion rate). Participants were paid \$1 for this experiment, which took roughly 10 min (\$6/hr).

**Materials and procedure.** For this experiment, we used the exact stimuli from Experiment 1. We increased the tempo of the stimuli by 23%, the maximum speed-up at which the stimuli still sounded natural. We used the free software Audacity (<https://www.audacityteam.org/>), so that the speed of the stimuli changed, but the pitch and formants remained unchanged. Because static spectral cues are highly predictive of the /s/ versus /ʃ/ contrast (e.g., McMurray et al., 2012, Table 3), this procedure is unlikely to change the perceived shift of our exposure tokens. Because these cues are duration invariant, we also do not expect that the increase in speech rate during exposure affects the perception of the test stimuli (that had the same speech rate as in Experiments 1 and 2).

The procedure of Experiment 3a was identical to that of Experiment 1, except that we added an additional prompt for participants in the Tongue Twister Context condition. This prompt emphasized the tongue twister as a plausible cause for the atypical pronunciations. We did so because Experiment 1 had not found an effect of the Tongue Twister Context on blocking the perceptual recalibration effect. Participants in the Tongue Twister Context were shown the following prompt:

A number of the phrases that the speaker was asked to say are difficult tongue twisters. You might notice that the speaker occasionally mispronounces certain words slightly because of this. Do not worry about the mispronunciations. Just transcribe the words as best as you can.

Participants in the Non-Tongue Twister Context were not shown the prompt. The rest of the procedure was identical to that of Experiment 1. Transcription accuracy (84.8%, *SD* = 10.1%) was somewhat lower than in Experiment 1, likely because of the increased speech rate in Experiment 3a.

### Results

To assess changes in the categorization boundary, we conducted the same analysis as in Experiments 1 and 2. We used mixed logit regression to predict /ʃ/ responses from Label (sum-coded: *f*-Label = 1 vs. S-Label = -1), Context (sum-coded: Non-Tongue Twister = 1 vs. Tongue Twister = -1), Trial Bin (coded continuously with the first trial bin as 0), and their interactions (see Table 9). For Experiment 3a, the analysis converged with by-participant intercepts and slopes for Trial Bin.

At the beginning of the test block, participants in the *f*-Label condition provided more /ʃ/ responses than those in the S-Label condition ( $\hat{\beta} = 0.63, z = 4.92, p < .001$ ). Furthermore, the total number of /ʃ/ responses tended to increase over trial bins ( $\hat{\beta} = 0.08, z = 3.1, p < .002$ ), and that this differed between Label conditions, in a way consistent with convergence toward 50/50: participants in the *f*-Label condition tended to provide fewer /ʃ/ responses in later trials bins, compared with those in the S-Label condition ( $\hat{\beta} = -0.15, z = -5.79, p < .001$ ). Critically, however, we did not identify a significant effect of Context ( $p = .54$ ) or

Table 9  
*Mixed Logit Regression Predicting Proportion of /ʃ/ Responses From Label, Condition, and Their Interaction (Experiment 3a)*

Predictors	Parameter estimates		Significance test	
	Coef ( $\hat{\beta}$ )	<i>SE</i>	<i>z</i>	<i>p</i>
(Intercept)	-.47	.13	-3.69	<b>&lt;.001</b>
Label ( <i>f</i> vs. S)	.63	.13	4.92	<b>&lt;.001</b>
Context (NonTT vs. TT)	.08	.13	.61	.54
TrialBin (first bin = 0)	.08	.03	3.1	<b>&lt;.002</b>
Label:Context	.08	.13	.67	.50
Label:TrialBin	-.15	.03	-5.79	<b>&lt;.001</b>
Context:TrialBin	-.04	.03	-1.52	.13
Label:Context:TrialBin	-.01	.03	-.40	.69

*Note.* Coding: Label (sum coded: *f*-Label = 1 vs. S-Label = -1), context (NonTT = 1 vs. TT = -1), trial bin (first bin = 0). Rows that are critical to our analysis are highlighted in grey. Significant effects are shown in bold, marginal effects in italics.

interaction between Context and Label ( $p = .50$ ). This suggests that the categorization of stimuli during the test block was not strongly affected by whether the shifted stimuli were presented in a Tongue Twister Context or Non-Tongue Twister Context. It is worth pointing out though that the interaction is numerically in the predicted direction. This is also visible in Figure 6.

Additional planned analyses reported in the online supplemental data found that (a) the magnitude of perceptual recalibration in Experiment 3a was identical to that of Experiment 1 and (b) Experiment 3a again only finds perceptual recalibration in the S-Label condition (compared with the unshifted baseline from Experiment 2). These results also suggest that the increased speech rate did not affect perceptual recalibration.

### Experiment 3b

In Experiment 3b, we further explore whether explicitly provided information about the talker can affect adaptation. We attempt to block perceptual recalibration by providing participants with an alternate reason why the talker might sound atypical. Specifically, we test whether instructions that talker in the experiment was intoxicated during the exposure block, but not during the test block, reduce or block the perceptual recalibration effect. We chose to use this alternate cause for two reasons. First, when intoxicated, speech errors become more common (Chin & Pisoni, 1997; Cutler & Henton, 2004). Second, the specific significant shift that we used to observe perceptual recalibration (/s/ shifting toward /f/) has been documented as one effect of intoxication on

speech production (Chin & Pisoni, 1997; Heigl, 2018). Both of these factors combined make it plausible that intoxication may provide listeners with a plausible cause for the atypical pronunciation that they hear.

### Method

**Participants.** We recruited 180 participants on MTurk, for a target of 40 participants in each of four conditions (S-Label/*f*-Label  $\times$  Intoxicated/Sober). Three participants were excluded for not correctly identifying the speaker as female, seven participants for not wearing headphones, and one for inverted categorization functions. We included one additional catch question in our post-experiment questionnaire to verify that participants were reading directions and were aware of when the talker was intoxicated or not (explained in Materials below). Nine additional participants were excluded for providing the incorrect response to this catch question (overall exclusion rate: 11.1%). Participants were paid \$1 for this experiment, which took roughly 10 min (\$6/hr).

**Materials and procedure.** The stimuli used during the exposure block were identical to the stimuli used in the Non-Tongue Twister condition of Experiment 1. As in Experiment 1, for half of participants, the atypical, shifted pronunciations occurred in words containing /s/ (S-Label condition), and for the other half of participants, the atypical, shifted pronunciations occurred in words containing /f/ (*f*-Label condition).

In both the Intoxicated and Sober conditions, participants were told that the purpose of the experiment was to understand how

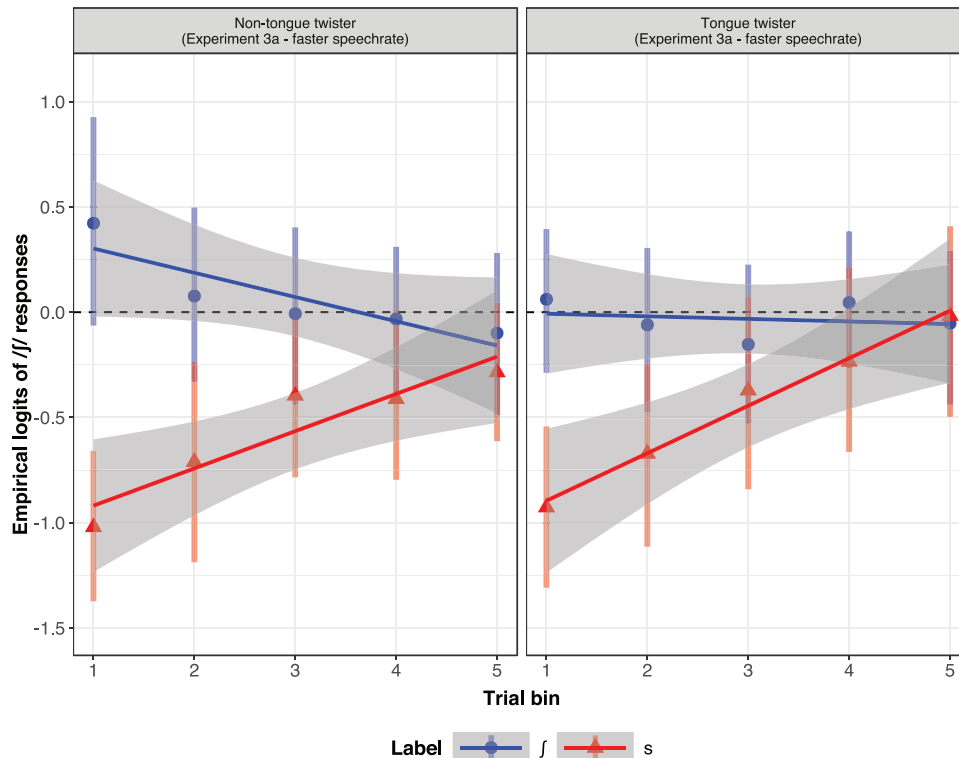


Figure 6. Empirical logits of /f/ responses as a function of Trial Bin (Experiment 3a). For further information, see caption of Figure 4. See the online article for the color version of this figure.

“people understand speech from speakers who are either intoxicated or not.” These groups only differed by the instructions that they saw preceding the exposure block. In the Sober condition, participants were told that they would hear words produced by “a speaker who is NOT intoxicated.” In the Intoxicated condition, participants were told that they would hear words produced by “an intoxicated speaker who had just drunk several cans of beer.”

Crucially, in both conditions, preceding the test block, participants were told that the same talker recorded additional words one week later. They were told that during this recording session, the speaker “reported that she had NOT drunk any beer, wine, or other alcoholic beverage in the past three days” and that we “confirmed this by testing her blood alcohol content (BAC = 0.00).” The rationale was that participants who were told that the talker was intoxicated would attribute the atypical pronunciations to her temporary, intoxicated state, and that their responses during the test block would, therefore, not show an effect of perceptual recalibration.

In the postexperiment questionnaire, we added an extra question to verify that participants read the critical prompt regarding the state of intoxication of the talker. Specifically, we asked the following:

The instructions that you read told you when the speaker you heard was intoxicated or not intoxicated. Please select the statement that best describes what the instructions told you. Reminder: the first section was where the speaker produced four word phrases, and the second section was where they produced asi/ashi words.

The possible responses were:

1. First section: Intoxicated. Second section: Not intoxicated.
2. First section: Not intoxicated. Second section: Intoxicated.
3. Both sections: Intoxicated.
4. Both sections: Not Intoxicated.

The correct response for participants in the Intoxicated condition was (1) and the correct response for participants in the Sober condition was (4). As reported above, we excluded participants when they answered this critical question incorrectly. Transcription accuracy (89.6%,  $SD = 6.8\%$ ) was similar to Experiment 1, which is expected given that the stimuli in Experiment 3b are identical to those of the Non-Tongue Twister condition in Experiment 1.

## Results

We again used mixed logit regression to predict /f/ responses from Label (sum-coded: f-Label = 1 vs. S-Label = -1), Context (sum-coded: Sober = 1 vs. Intoxicated = -1), Trial Bin (coded continuously with the first trial bin as 0), and their interactions (see Table 10).

At the beginning of the test block, participants in the f-Label condition provided more /f/ responses than those in the S-Label condition ( $\hat{\beta} = 0.58$ ,  $z = 5.13$ ,  $p < .001$ ). Critically, however, we did not identify a significant effect of Context (Sober vs. Intoxi-

Table 10  
Mixed Logit Regression Predicting Proportion of /f/ Responses From Label, Condition, and Their Interaction (Experiment 3b)

Predictors	Parameter estimates		Significance test	
	Coef ( $\hat{\beta}$ )	SE	z	p
(Intercept)	-.35	.11	-3.09	<b>&lt;.001</b>
Label (f vs. S)	.58	.11	5.13	<b>&lt;.001</b>
Context (sober vs. intoxicated)	.03	.11	.25	.8
TrialBin (first bin = 0)	.02	.02	.93	.35
Label:Context	.09	.11	.77	.44
Label:TrialBin	-.08	.02	-3.81	<b>&lt;.001</b>
Context:TrialBin	.01	.02	.32	.75
Label:Context:TrialBin	-.02	.02	-.88	.38

Note. Coding: Label (sum coded: f-Label = 1 vs. S-Label = -1), context (sober = 1 vs. intoxicated = -1), trial bin (first bin = 0). Rows that are critical to our analysis are highlighted in grey. Significant effects are shown in bold.

cated:  $p = .8$ ) or interaction between Context and Label ( $p = .44$ ). It is worth pointing out though that the interaction is again numerically in the predicted direction, as it was in Experiment 3a. This is also visible in Figure 7. Further simple effect comparison against Experiment 2 confirmed that, again, the perceptual recalibration we found in Experiment 3b was driven by only the S-Label condition (analysis not reported here).

Additional planned analyses reported in the online supplemental data found that (a) the magnitude of perceptual recalibration in Experiment 3b was identical to that of Experiment 1 and (b) Experiment 3a again only finds perceptual recalibration in the S-Label condition (compared with the unshifted baseline from Experiment 2). This, too, suggests that the manipulation in Experiment 3b did not affect perceptual recalibration.

## Discussion

The results of Experiments 3a and 3b are interesting in light of some experiments that have found causal attribution effects of information provided via explicit instructions on aspects of language processing, other than speech perception (e.g., Arnold et al., 2007; Grodner & Sedivy, 2011; Niedzielski, 1999). For example, Arnold et al. (2007) reported that explicitly telling participants that a talker had object agnosia lead to differences in participant expectations during reference comprehension. Other experiments, however, have found little (Pogue, Kurumada, & Tanenhaus, 2016) or no role of explicit instructions (Dix et al., 2018). We discuss these and other studies in the context of the current experiments in more detail in the General Discussion.

In Experiments 3a and 3b, we again observe robust perceptual recalibration effects for exposure to shifted pronunciations of /s/. Furthermore, we again fail to observe significant blocking of perceptual recalibration in the presence of an incidental cause for the shifted pronunciations. That is, unlike visual evidence of a pen in the mouth during exposure, none of the incidental causes explored in Experiments 1, 3a, and 3b seems to prevent perceptual recalibration. The findings of Experiments 1–3 are problematic for theories that attribute the effect



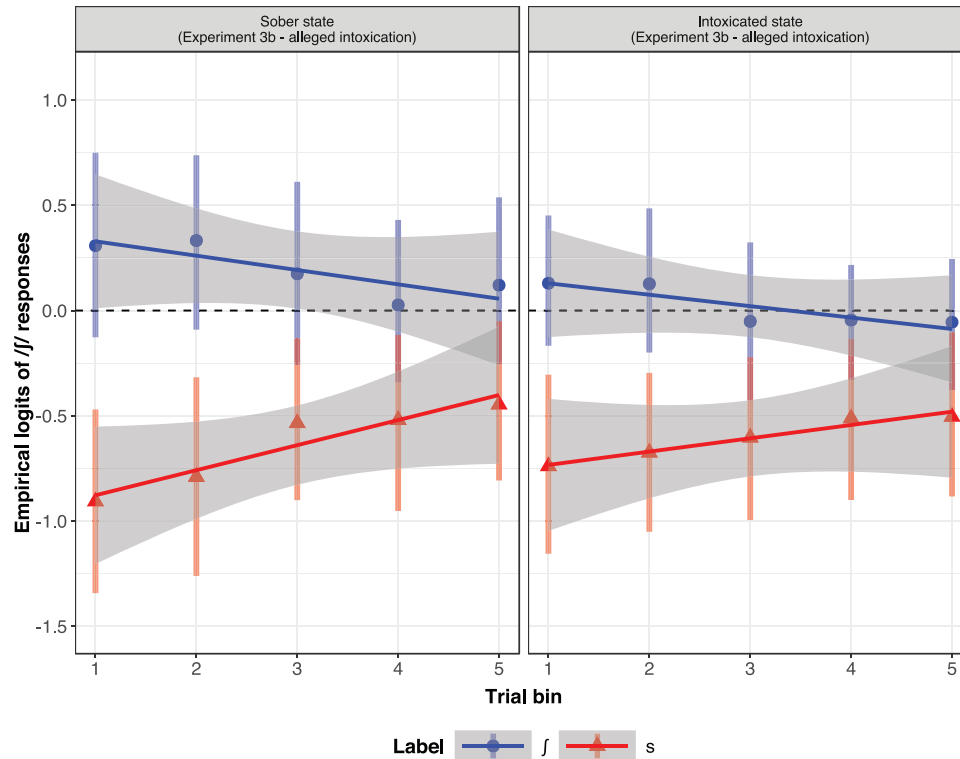


Figure 7. Empirical logits of /f/ responses as a function of Trial Bin (Experiment 3b). For further information, see caption of Figure 4. See the online article for the color version of this figure.

of the pen in the mouth to inferences about causes during speech adaptation (Liu & Jaeger, 2018), although both experiments exhibit trends in the predicted direction. One possibility that explains why we may have failed to find blocking of adaptation is that participants do not perceive the tongue twisters, even with the explicit instructions, as plausible tongue twisters. We explore this possibility in Experiments 4a and 4b.

### Experiments 4a and 4b

Experiment 4 assesses the possibility that participants in Experiments 1 and 3a considered it implausible that the shifted pronunciations in the Tongue Twister Context were because of incidental speech errors, rather than being characteristic of the talker. Two considerations guide the design of Experiment 4.

First, speech errors are very rare in everyday speech production. Hearing multiple speech errors—all of them on the same type of sound—make it less likely that the mispronunciation is not characteristic of the talker, even when those mispronunciations occur in a tongue twister. Specifically, Experiments 1 and 3a exposed participants to eight different speech errors, all involving the same sound (either /s/ or /f/). Given that speech errors only occur in tongue twister contexts about 8–17% of the time (Choe & Redford, 2012; Motley & Baars, 1976), this high incidence of errors could have led participants to infer that the typical pronunciations are characteristic of how the talker

typically sounds. Second, it is possible that some of our tongue twister contexts are perceived to be more likely to cause speech errors than others. The less plausible a tongue twister context is perceived to be, the more likely listeners should be to attribute the shifted pronunciation to the talker rather than the context. Either of these two possibilities could explain the failure to observe blocking of perceptual recalibration in Experiments 1 and 3a.

In Experiment 4, we attempt to remedy both of these concerns. We identify the top four plausible tongue twisters in Experiment 4a, cutting down the number of tongue twisters we use from eight to four. In Experiment 4b, we first validate that exposure to only these four items in a Non-Tongue Twister Context results in perceptual recalibration (it does), and then test whether exposure to the same four items in a Tongue Twister Context blocks perceptual recalibration.

### Experiment 4a

#### Method

**Participants.** There were 90 participants who participated in our experiment to achieve a target of 20 participants for each of the four between-participants conditions (S/f-Label crossed with the Tongue Twister/Non-Tongue Twister Context). They were paid \$0.50 for this Experiment, which took about 5 min (\$6/hr). Four participants were excluded for not answering the catch question

correctly; six were excluded for reporting that they did not wear headphones.

**Materials and procedure.** In this experiment, participants listened to stimuli one at a time and rated them on a Likert scale from 1 (*not at all like a tongue twister*) to 7 (*definitely a tongue twister*). Participants were presented with the 30% sped-up stimuli from Experiment 3a (using the same lists), with the addition of three control tongue twisters. The three control tongue twisters were taken from well-known tongue twisters, and were adjusted to roughly match the structure of the other phrases (disyllable words, repeated words):

- Betty Botter Butter Betty
- Peter Piper Pepper Peter
- Soldier Shoulder Soldier Shoulder

Each participant provided 27 total judgments. Following the experiment, participants completed the same postexperiment questionnaire as in Experiments 1–3.

## Results

First, we wished to assess whether our tongue twisters were perceived as more tongue-twister-like than the filler stimuli. We compared the ratings of our tongue twister stimuli to both the filler stimuli and attested tongue twisters. Figure 8 shows the mean ratings for attested tongue twisters, our tongue twisters, nontongue twister stimuli, and fillers without any /s/ or /f/. To remove individual variability in how participants used the rating scale, we first standardized ratings by participant for plotting. Linear mixed-effects analyses of the unstandardized ratings are reported in the online supplemental data, and confirmed what is visible in Figure 8. Tongue twisters were rated as more tongue-twister-like than the filler stimuli (as intended), but less tongue-twister-like than attested tongue twisters. We

postpone discussion of possible reasons for this until after Experiment 4b. We found no differences in ratings between the S- and f-Label condition for any of the different stimuli types, including the Tongue Twisters we created for our experiments ( $p > .3$ ; S-Label: mean rating = 4.7 ( $SD = 1.0$ ); f-Label: mean rating = 4.5 ( $SD = 0.7$ )). This suggests that the Tongue Twister Context was not viewed as more plausible for one Label condition compared with the other.

Second, we wished to assess whether certain tongue twisters within our stimuli were perceived as more tongue-twister-like than others. For each Label condition, we computed the average ratings for each of the eight tongue twister phrases. These averaged between 3.5 and 5.4. We then selected the four tongue twisters out of these eight with the highest tongue twister ratings in both the S-Label and f-Label conditions. We use these tongue twisters for a shortened version of Experiment 1 in Experiment 4b.

## Experiment 4b

### Method

**Participants.** There were 179 participants who participated in this experiment, for achieve the targeted 40 participants in each of four conditions (S-Label/f-Label  $\times$  Tongue Twister/Non-Tongue Twister Context). Fourteen participants were excluded for providing an incorrect answer to the catch question, three participants were excluded for an inverted category boundary, and two participants were excluded for not wearing headphones (10.6% overall exclusion rate). Participants were paid \$0.70 for this experiment, which took about 7 min (\$6/hr).

**Materials and procedure.** The materials that we used for this experiment were a subset of the materials used in Experiment 3a

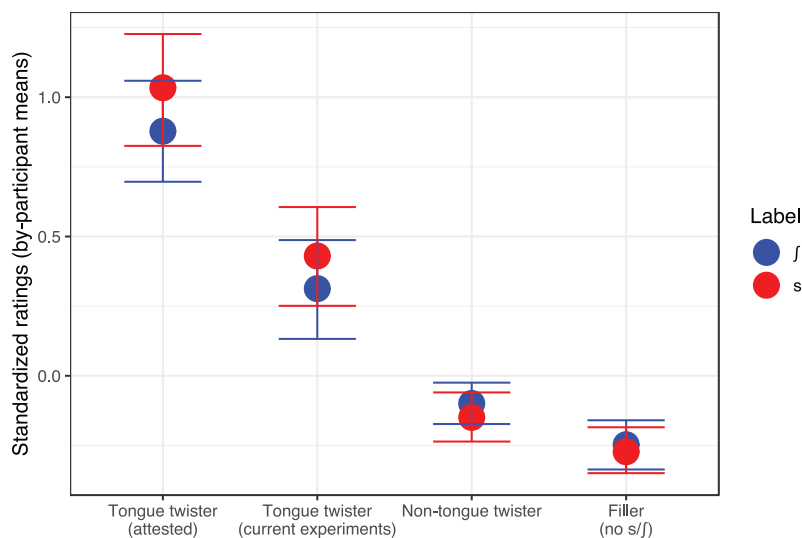


Figure 8. Mean ratings for each stimulus type (Experiment 4a). For plotting, responses were standardized within participants before taking by-participant means (high values indicate “more tongue twister-like”). Error bars show 95% confidence intervals obtained via nonparametric bootstrap over the by-participant means. See the online article for the color version of this figure.

(i.e., the stimuli for which speech rate was increased by 30% compared with Experiment 1). For each Label condition, these consisted of half of the critical phrases (four instead of eight), as well as half of the filler phrases (eight instead of 16). These are the items marked with an asterisk (\*) in Tables 1 and 2. Transcription accuracy (82.8%,  $SD = 11.4%$ ) was about 6% lower than in Experiment 1 and about 2% lower than in Experiment 3a.

## Results

As in all previous experiments, we performed mixed logit regression to predict /f/ responses from Label (sum-coded: f-Label = 1 vs. S-Label = -1), Context (sum-coded: Non-Tongue Twister = 1 vs. Tongue Twister = -1), Trial Bin (coded continuously with the first trial bin as 0), and their interactions. This is shown in Table 11 and visualized in Figure 9.

The results are visualized in Figure 9. At the beginning of the test block, participants in the f-Label condition provided more /f/ responses than those in the S-Label condition ( $\beta = 0.36$ ,  $z = 4.48$ ,  $p < .001$ ). This shows that the four shifted exposure items were sufficient to elicit perceptual recalibration. There was no main effect of Context ( $p = .95$ ) and, critically, we did not identify a significant interaction between Context and Label ( $p = .19$ ), though the interaction again trends in the predicted direction (as also evident in Figure 9).<sup>4</sup>

Additional planned analyses reported in the online supplemental data found that the magnitude of perceptual recalibration in Experiment 4b (4 shifted sounds, 0.72 log-odds perceptual recalibration) was numerically, but not significantly, smaller than in Experiment 3a (eight shifted sounds, 1.18 log-odds). This is expected given that the two experiments differed in the number of shifted sounds.

## Discussion

Experiment 4 reduced the number of critical items with atypical, shifted pronunciations to reduce the probability that listeners would infer that the shifted sound is characteristic of the talker. We halved the number of critical tongue twister items and fillers, and used only the four most plausible tongue twister contexts. We

again identified evidence for perceptual recalibration in both the Tongue Twister and Non-Tongue Twister Contexts. We observed numerically, but not significantly, weaker perceptual recalibration (the simple effect of Label condition) in the Tongue Twister Context than in the Non-Tongue Twister Context, in line with the hypothesis that participants might attribute the talker's shifted pronunciation to the context.

It is possible that we were unable to detect this effect because the tongue twisters we created were not perceived to be sufficiently plausible to elicit shifted pronunciations. While our Tongue Twister contexts were rated as more tongue twister-like than our Non-Tongue Twister context, they received lower ratings compared with attested tongue twisters. This might make it less likely that participants attribute the atypical pronunciations to the tongue twister, instead attributing it to the talker. This in turn would explain the lack of a significant interaction of the Label and Context conditions. We return to this possibility in the general discussion. First, we present one final experiment aimed at increasing the probability that listeners interpret the shifted pronunciations in Tongue Twister contexts as incidental errors, rather than characteristic of the talker.

## Experiment 5

In Experiment 5, we provide additional bottom-up evidence for the Tongue Twister Context, to increase the plausibility of our tongue twister phrases. We conducted an informal review of speech errors in tongue twisters on Youtube.com, and observed that naturally occurring tongue twisters often contain additional evidence that a talker experienced a production difficulty. Indeed, self-corrections have been observed in more than 50% of all speech errors (Nooteboom, 1980), and self-monitoring mechanisms appear in the majority of models of speech production (see Postma, 2000 for a review). We incorporate these properties of speech errors into the stimuli for our Experiment 5.

We exposed participants to the same stimuli as in Experiment 4b, except that we edited the stimuli to add auditory evidence of a repair or audible frustration because of production difficulty. Specifically, we created two new Tongue Twister Context conditions. In the first new condition, the talker makes a stutter during the atypical pronunciation following the first syllable, and then attempts to repair their error by repeating the word (Difficulty During Context). In our stimuli, this repair always resulted in a second atypical pronunciation of the same word. This design decision was made so as to avoid presenting both typical and atypical pronunciations of the same sound, which would be a deviation from perceptual recalibration paradigms. We note, however, that this context condition may unintentionally reinforce the possibility that the atypical pronunciation reflects how the talker typically sounds, as the talker's repair still contains an atypical pronunciation.

The prediction for the Difficulty During Context depends on how participants (on average) interpret this condition. If participants interpret this context as an unsuccessful repair, and, thus, as

Table 11  
Mixed Logit Regression Predicting Proportion of /f/ Responses From Label, Condition, and Their Interaction (Experiment 4b)

Predictors	Parameter estimates		Significance test	
	Coef ( $\hat{\beta}$ )	SE	z	p
(Intercept)	-.02	.08	-.27	.79
Label (f vs. S)	.36	.08	4.48	<.001
Context (NonTT vs. TT)	-.01	.08	-.07	.95
TrialBin (first bin = 0)	.02	.02	1.08	.28
Label:Context	.11	.08	1.32	.19
Label:TrialBin	-.09	.02	-4.33	<.001
Context:TrialBin	-.02	.02	-.82	.41
Label:Context:TrialBin	-.05	.02	-2.36	<.05

Note. Coding: Label (sum coded: f-Label = 1 vs. S-Label = -1), context (NotTT = 1 vs. TT = -1), trial bin (first bin = 0). Rows that are critical to our analysis are highlighted in grey. Significant effects are shown in bold.

<sup>4</sup> We note that our approach to estimate effects in the first trial bin seems to over-estimate perceptual recalibration in the Tongue-Twister condition, compared with the Non-Tongue Twister condition. We return to this in the discussion.

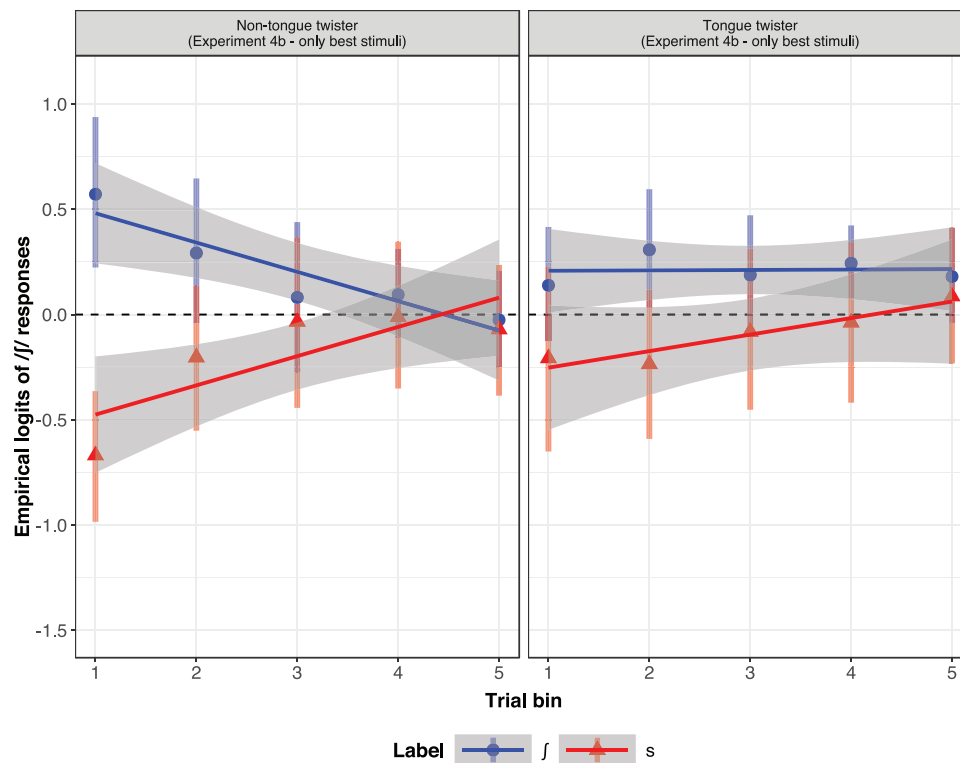


Figure 9. Empirical logits of /f/ responses as a function of Trial Bin (Experiment 4b). For further information, see caption of Figure 4. See the online article for the color version of this figure.

evidence that the talker considered the atypical pronunciation to deviate from the intended pronunciation, perceptual recalibration should be reduced or blocked. If, on the other hand, participants take the repeated atypical pronunciation as further evidence that the atypical pronunciation is characteristic of the talker, we should see as much or more perceptual recalibration in this context condition, as in the Non-Tongue Twister condition.

The second new context condition of Experiment 5 avoids this problem, leading to clearer predictions. In this condition, the talker emits a sound of frustration *following* the production of the atypical pronunciation (Difficulty After Context). This provides the listener with evidence that the talker is aware of the production error that they made, and that they found it to be deviating from their own internal criterion of how that word should sound. We note that the Difficulty After Context also corresponds most closely to what we observed in the majority of speech errors in tongue twisters on Youtube.com. The predictions for the Difficulty After Context are clear: if causal inferences attribute the atypical pronunciation to a speech error, perceptual recalibration should be reduced or blocked in this condition, compared with the Non-Tongue Twister Context.

## Method

**Participants.** There were 167 participants who participated in this Experiment 5, for a target of 40 participants in each Label and Context condition (S-Label/f-Label  $\times$  During/After). Three par-

ticipants were excluded for providing an incorrect answer to the catch question and four participants were excluded for an inverted category boundary (4.2% overall exclusion rate). Participants were paid \$0.70 for this 7 min long experiment (\$6/hr).

**Materials and procedure.** The materials for this experiment were identical to the ones used in the Tongue Twister Context condition in Experiment 4b, with slight modifications to the critical stimuli. Namely, in the During Context, for the eight critical phrases containing an atypical pronunciation, we inserted a sign of production difficulty before the second syllable of the atypical pronunciation, followed by a repair. For example, for the phrase “passion mansion pa?shive passion” was produced as “passion massion pa [stutter] pa?sfive passion.” In the After Context, after the atypical pronunciation, we inserted a sigh to signal frustration. For example, for the phrase “passion mansion pa?sfive passion” was produced as “passion massion pa?sfive [ugh] passion.” A separate norming experiment (reported as Experiment 5b in the online supplemental data) verified that participants indeed perceived that the talker had more difficulty with the phrases containing overt signs of production difficulties (Experiment 5), compared with those without (Experiments 1–4).

With 80.4% ( $SD = 10.2\%$ ), transcription accuracy in Experiment 5 was the lowest of all experiments, possibly because the additional difficulty of transcribing phrases with overt signs of production difficulty. For example, participants might have been unsure whether or not to transcribe the word containing the production difficulty or repair.



## Results

Our goal is to assess whether either the insertion of a sign of production difficulty coupled with a repair (During Context) or a sign of production difficulty following an atypical pronunciation (After Context) in the Tongue Twister Context may result in a reduction or blocking of the perceptual recalibration effect. To this end, we compare these two conditions of Experiment 5 to the Non-Tongue Twister condition of Experiment 4b. We wished to assess whether either the insertion of a sign of production difficulty coupled with a repair (During Context) or a sign of production difficulty following an atypical pronunciation (After Context) in the Tongue Twister Context may result in a reduction or blocking of the perceptual recalibration effect identified in the Non-Tongue Twister Context.

We performed mixed logit regression to predict /f/ responses from Label (sum-coded: f-Label = 1 vs. S-Label = -1), Context (treatment coded, with the Non-Tongue Twister Context as the comparison level), Trial Bin (coded continuously with the first trial bin as 0), and their interactions. This is shown in Table 12 and visualized in Figure 10.

A significant interaction between Context and Label for either of the comparisons of the Tongue Twister Contexts against the Non-Tongue Twister Context (During Context vs. Non-Tongue Twister Context, or After Context vs. Non-Tongue Twister Context) would suggest that the Tongue Twister Context, when combined with a sign of production difficulty, resulted in a difference in participants' categorization boundaries during the test block, compared with the Non-Tongue Twister Context.

Though we found that participants in the During Context provided overall significantly fewer /f/ responses than those in the Non-Tongue Twister Context ( $\hat{\beta} = -0.37, z = -2.0, p < .05$ ), we failed to identify any significant interaction between Label and either Context comparison ( $ps > .41$ ). Simple effects analysis revealed that the effect of Label was significant for all three

contexts ( $ps < .05$ ), though numerically smaller in the After Context (difference between the f-Label and S-Label condition = 0.66 in log-odds) compared with the Non-Tongue Twister Context (0.96 in log-odds) and the During Context (1.26 in log-odds). This is visualized in Figure 10 (see differences in Trial Bin 1).

In summary, we again find a robust effect of perceptual recalibration, and no significant effects that would indicate that listeners take into account incidental causes. We note, however, that the relative size of the perceptual recalibration effects is in line with a possibility we raised above: participants might have interpreted only the Difficulty After Context as good evidence that the talker recognized the atypical pronunciation as unintended (and, thus, not characteristic of her speech); unintentionally, the design of our Difficulty During Context—that involved repetition of the atypical pronunciation in the repair—might have reinforced, rather than weakened, participants' belief that the atypical pronunciation is characteristic of the talker. This would explain the numerical pattern we observed in Experiment 5. We return to this possibility below.

## General Discussion

In five experiments, we explore the role of inferences about alternative causes during speech perception. We identify robust perceptual recalibration following exposure to as few as four and eight shifted pronunciations embedded within four-word phrases. Recalibration was observed despite the fact that critical target words with atypical pronunciations only account for less than 10% of all words heard during exposure (see also Kraljic & Samuel, 2005; Kraljic et al., 2008). This reliably replicates perceptual recalibration in a web-based paradigm, despite variability in the audio equipment across participants (see also Kleinschmidt & Jaeger, 2012; Liu & Jaeger, 2018). Additionally, we replicate the effect that learning from exposure is unlearned during test, because of exposure to a uniform distribution of sounds during test (Liu & Jaeger, 2018). This confirms that the common practice of reporting recalibration averaged across the entire test phase systematically underestimates the adaptivity of the perceptual system: at the beginning of the test phase, perceptual recalibration is often twice as large as when averaged across all block. The analysis we used throughout the present study takes this into account.

The goal of the present study was to identify whether perceptual recalibration is affected by the presence of an alternative cause for the atypical pronunciations. Previous studies found that exposure to an atypical pronunciation paired with a video of the talker with a pen in her mouth resulted in a complete blocking of the adaptation effect for at least one of the exposure (Label) conditions (Kraljic & Samuel, 2011; Kraljic et al., 2008; Liu & Jaeger, 2018). One explanation for this is that listeners attribute the atypical pronunciation to the pen. This would block perceptual learning, either because listeners do not store the atypical pronunciations as part of their talker-specific experience (because they do not attribute the atypicality of the pronunciation to the talker), or because listeners store the atypical pronunciations but do so together with the contextual information that a pen was in the speaker's mouth. According to the latter explanation (proposed in Kraljic & Samuel, 2011), perceptual recalibration is blocked because no pen is present during the test trials (unlike exposure trials, test trials in were auditory only) so that listeners might not consider the input they experienced during exposure as relevant to the categorization of

Table 12  
Mixed Logit Regression Predicting Proportion of /f/ Responses From Label, Condition, and Their Interaction (Experiment 5)

Predictors	Parameter estimates		Significance test	
	Coef ( $\hat{\beta}$ )	SE	z	p
(Intercept)	-.03	.13	-.23	.82
Label (f vs. S)	.48	.13	3.69	<b>&lt;.001</b>
Context1 (during vs. NonTT)	-.37	.19	-1.98	<b>&lt;.05</b>
Context2 (after vs. NonTT)	-.27	.19	-1.46	.15
TrialBin (first bin = 0)	.01	.03	.19	.85
Label:Context1	.15	.19	.78	.43
Label:Context2	-.15	.19	-.82	.41
Label:TrialBin	-.14	.03	-4.76	<b>&lt;.001</b>
Context1:TrialBin	.06	.04	1.38	.17
Context2:TrialBin	.01	.04	.18	.86
Label:Context1:TrialBin	-.02	.04	-.48	.63
Label:Context2:TrialBin	.07	.04	1.81	.07

Note. Coding: Label (sum coded: f-Label = 1 vs. S-Label = -1), context (treatment coded, with the Non-Tongue Twister Context as the comparison level), trial bin (first bin = 0). Rows that are critical to our analysis are highlighted in grey. Significant effects are shown in bold, marginal effects in italics.

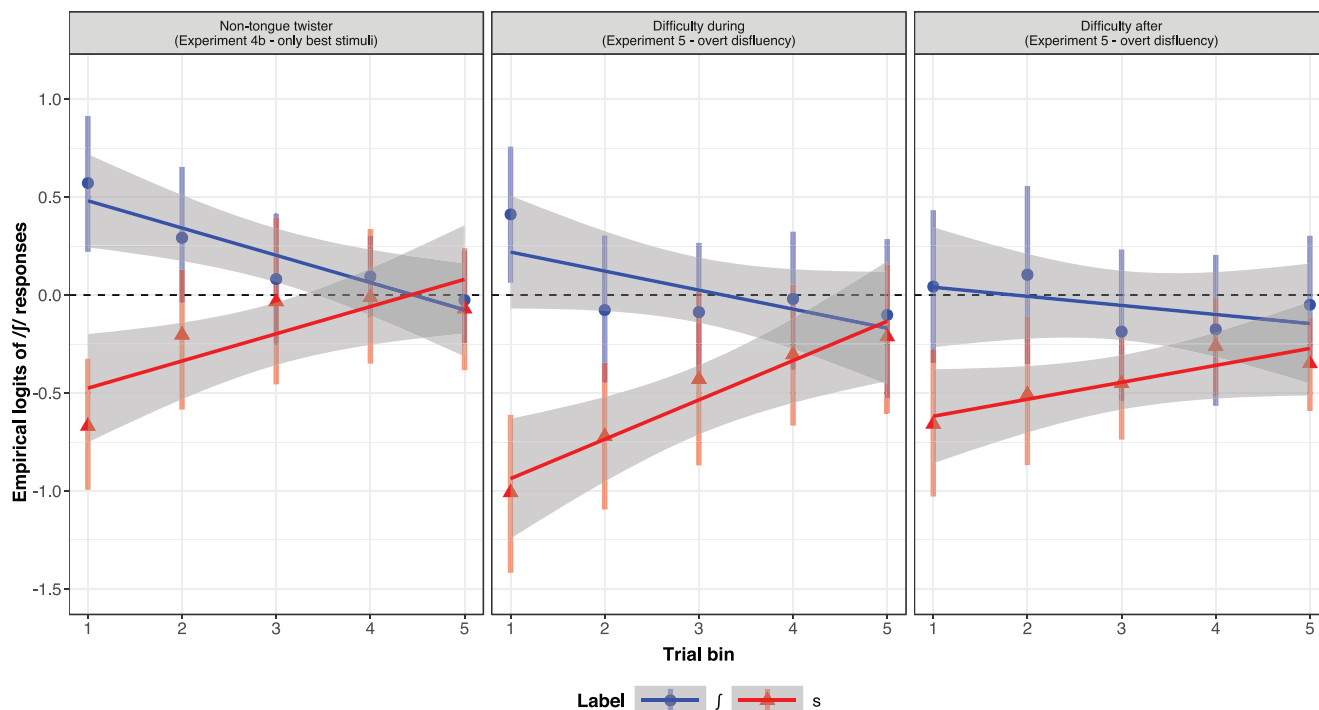


Figure 10. Empirical logits of /f/ responses as a function of Trial Bin (Experiment 5). For further information, see caption of Figure 4. See the online article for the color version of this figure.

the test stimuli. Either of these two explanations attributes pen-in-the-mouth effect to causal inferences—either during the storage of previously experienced input (to determine what constitutes *relevant* context) or during its retrieval (e.g., to determine what previous experience is relevant to the processing of the current input; see discussion in Liu & Jaeger, 2018).

In contrast to the pen-in-the-mouth effect, we do not find significant reduction of perceptual recalibration for any of the incidental causes we explored. It is, however, noteworthy that the interaction of Context and Label condition went in the expected direction in five out of six between-participants comparisons (in Experiments 3a, 3b, 4b, and in the Difficulty After condition in Experiment 5). This consistency in the nonsignificant trends is quite unexpected by chance alone (Wilcoxon signed-ranks test with continuity correction:  $V = 17.5$ ,  $p = .13$ ). Prompted by reviews, we conducted two sets of post hoc tests. Before we discuss our results further, we briefly summarize these tests (for details, see the additional online data at <https://osf.io/ungba/>).

### Summary of Post Hoc Tests Prompted by Reviews

The first post hoc test pooled all experiments together to assess the effect of Context (sum-coded: Non-Tongue Twister or Sober = 1 vs. Tongue Twister or Intoxicated = -1). This mixed logit regression was otherwise identical to the analyses of all individual experiments, including all other predictors, their coding, and the random effects. The critical interaction between Label condition and Context went in the expected direction, but was not significant ( $\beta = 0.06$ ,  $z = 1.3$ ,  $p = .19$ ). We also repeated this analysis with a Bayesian framework to obtain a well-formed, and more intuitive,

measure of evidentiary support (Wagenmakers, 2007). This analysis compared the relative probability of the hypothesis that incidental causes reduce the magnitude of perceptual recalibration against the hypothesis that they increase the magnitude of perceptual recalibration. In line with the numerical trend we observe, the analysis estimated the posterior probability of the former hypothesis to be 89% ( $BF = 8.0$ ; for details, see additional online data).

This first post hoc test ignored that we *expect* stronger effects in the latter experiments (for all the same reasons that motivated these experiments, on which we further elaborate below). The second set of post hoc tests reanalyzes the data from all experiments separately. These additional tests also address another potential shortcoming of the analyses presented above: As one reviewer pointed out, our procedure of estimating effects of the Label condition in the first trial bin sometimes *over*-estimates perceptual recalibration in the Tongue Twister condition and *under*-estimates the perceptual recalibration in the Non-Tongue Twister Condition. This is visible, for example, in Figure 4 for Experiment 4b.

Thus, we repeated the same analysis reported above for all experiments, but over only the responses in the first trial bin (i.e., excluding all other data and excluding Trial Bin as a predictor). These analyses have less power as they are based on less data, but do not make the linearity assumption made in the main analyses reported above. This second set of post hoc analyses replicated the significant main effect of Label for all experiments for all experiments ( $ps < .004$ ). The critical interaction between Label and Context was significant for Experiment 4b ( $\beta = 0.21$ ,  $z = 2.31$ ,  $p < .02$ ), surviving Bonferroni correction ( $\alpha_{corrected} = 0.025$ ). Simple effects analysis revealed significant effects of the Label condition for the Non-Tongue Twister

condition ( $\hat{\beta} = 0.61, z = 4.77, p < .0001$ ), but not the Tongue Twister condition ( $\hat{\beta} = 0.19, z = 1.56, p < .12$ ). In all other experiments, the interaction between Context and Label was not significant ( $ps > .14$ ). In short, while the present studies return some evidence compatible with causal inference accounts, this support is weak—in contrast to previous studies with visually presented causes. All post hoc analyses are reported in full in the additional online data, along with the data from all experiments.

We close by discussing explanations of our results. We first discuss the possibility that causal inference during speech perception is limited to certain types of incidental causes. Then we discuss alternative interpretations of our findings, and raise design considerations for future work.

### Incidental Causes During Spoken Language Understanding

One explanation for the difference between the present findings and those of earlier work is that visual information—perhaps, in particular, visual information about articulation that occurs concurrently with the auditory input—has a privileged role during speech perception. Such visual information can influence phoneme perception and appears to be strongly integrated with auditory input (e.g., Tuomainen et al., 2005). This is demonstrated by the McGurk effect, in which an auditory /ba/ dubbed onto a video of a talker producing /ga/ results in a percept of /da/ (McGurk & MacDonald, 1976). The incidental causes used in the present study either were not presented visually (tongue twisters) or were presented visually, but did not constitute audio-visual speech percept (explicit instruction preceding exposure, e.g., when participants were told that the talker was intoxicated).

We know of no previous work that has directly addressed whether causal inference during perception are affected by incidental causes that are *not* presented visually and concurrently with the speech signal.<sup>5</sup> There are, however, two lines of research that are of relevance to this question.

First, there is the observation that listeners are generally capable of integrating evidence about a talker from a diverse array of sources—both visual and nonvisual during language comprehension—and regardless of whether the evidence is presented concurrently with the language input (e.g., Arnold et al., 2007; Dix et al., 2018; Grodner & Sedivy, 2011; Hay, Nolan, & Drager, 2006; Hay, Warren, & Drager, 2006; McGowan, 2015; Pogue et al., 2016). This includes hypothetical incidental causes that are indicated through explicit instructions. For example, Grodner and Sedivy (2011) provided listeners with instructions that a talker “had an impairment that caused language and social problems.” They found that listeners used these instructions to modulate their pragmatic processing of sentences from that (unreliable) talker (see also Dix et al., 2018). A similar effect has been observed for the processing of disfluencies (Arnold et al., 2007). Eye-tracking visual world experiments demonstrate that listeners are sensitive to the presence of disfluencies: following a disfluency (“Click on [pause] thee uh red . . .”), listeners anticipated references to unfamiliar objects with difficult names, as opposed to familiar object with simpler names (see also Arnold, Tanenhaus, Altmann, & Fagnano, 2004). In the absence of a disfluency, listeners’ eye-movements exhibited the opposite preference. This suggests that listeners take into account that disfluencies tend to precede

referential expressions that are associated with production difficulty. Crucially, when another group of listeners was told that the talker had a language impairment—a difficulty recognizing and naming objects—listeners’ eye-movements no longer exhibited sensitivity to disfluencies. Arnold and colleagues interpreted this blocking of the typical interpretation of disfluencies to inferences about alternative causes for disfluencies (in this case, the talker’s language impairment). These findings suggest that explicitly indicated incidental causes can affect some aspects of language processing. It is worth noting though that these studies have investigated higher-level aspects of language processing, rather than speech perception.

A second line of experiments demonstrates that explicit instructions can guide listeners’ expectations about how a talker will sound. For example, informing a listener that a talker is from a particular region can affect vowel perception (Niedzielski, 1999; Hay et al., 2006). Niedzielski found that listeners interpreted the same vowel sound differently depending on whether they were told that the talker was from Canada or Detroit. The dialects of two regions differ in how they tend to pronounce the same vowels, and listeners’ interpretation of the acoustic input reflected these differences. Expectations about talker identity and accent do not necessarily need to be initiated through explicit instructions. Similar effects on vowel perception have been identified when listeners are provided with answer-sheets labeled as “Australian” or “New Zealander” (Australian and New Zealand English differ in their vowel system; Hay et al., 2006), when a stuffed animal strongly associated with either Australia or New Zealand was displayed in the experiment room (Hay & Drager, 2010), and when listeners were provided with images of talkers of different ages/social classes that were associated with particular vowel variants (Hay et al., 2006). Beyond vowel perception, similar effects have been identified for talker intelligibility (e.g., McGowan, 2015).

Results like these leave open whether specifically causal inferences during perception can be affected by explicit instructions. They also leave open whether visual information that is presented concurrently with the speech signal have a special status specifically with regard to causal inferences. Results like the ones summarized here do, however, argue that speech perception in general can be affected by these others sources of information. This leads us to discuss alternative explanations next. We discuss possible explanations for the fact that all but one of the non-significant context effects in the present experiments trended in the direction expected under the causal inference account.

### Directions for Future Work: Increasing the Probability That Atypical Pronunciations Are Inferred to Reflect Incidental Speech Errors

Listeners can be highly attuned to the plausibility of different causes for observations they make in the speech input. For example, in the aforementioned study by Arnold et al. (2007), though the authors found an effect of explaining away on reference comprehension when listeners were informed a talker had object

<sup>5</sup> Some studies have investigated whether the type of task listeners are instructed to do during exposure affects perceptual recalibration (for an excellent review and references, see Drouin & Theodore, 2018). These studies do not manipulate incidental causes, but rather aim to manipulate the degree of attention to specific aspects of language processing.

agnosia, they found no such effect when listeners were provided with evidence that the talker was distracted by construction noise or beeps. It is possible that listeners found the construction noise or beeps implausible causes for the talker's disfluency.<sup>6</sup> A couple of other studies have found similar sensitivity to subtle changes in the presentation of incidental causes (compare Dix et al., 2018; Grodner & Sedivy, 2011).

The present experiments were designed to make our tongue twister contexts plausible causes for speech errors. We used tongue twister contexts that were modeled after attested tongue twisters. We used critical sounds that are known to frequently be the target of speech errors. We created shifted sounds that were meant to resemble graded, noncategorical speech errors has been observed in naturally occurring tongue twisters. We then selected those tongue twisters that were rated to be most plausible by another set of participants (Experiments 4 and 5).

However, despite these precautions, it is possible that participants did not perceive our manipulations as likely to explain the atypical pronunciations. We discuss three properties of our experiments that might have contributed to this, all three of which are related to the "plausibility" of our stimuli under the hypothesis that the atypical pronunciations resulted from incidental speech errors.

First, it is possible that our tongue twister context were perceived as not sufficiently likely to induce *any* type of speech error. Naturally occurring tongue twister errors typically involve repeated reiteration of a phrase (e.g., *passion mansion passive passion passion mansion passive passion . . .*; Wilshire, 1999). For reasons we discuss next, we did not incorporate this property of natural tongue twisters into the design of our study. The present experiments are the first to investigate perceptual recalibration in the context of tongue twisters. Thus, we aimed to keep our paradigm as comparable as possible to previous work on perceptual recalibration. Any further deviation from typical perceptual recalibration paradigms would have to be carefully piloted (as we did in Experiments 1 and 2). In particular, repetition of tongue twister contexts would require design decisions as to whether the critical sound (e.g., /f/) is *always* shifted or only sometimes.

Neither design decision is without potential problems. If the critical sound is always shifted, this results in a large number of shifted sounds, which makes it more likely that the atypical pronunciations is characteristic of the talker. This potential confound would counter the intended effect of the manipulation. If, on the other hand, only some of the instances of the critical sound are shifted, this would provide listeners with evidence that the talker does *not* always produce the shifted pronunciation. As exposure to some shifted and some normal pronunciation is likely to result in less perceptual recalibration, it would be important to also compare this hypothetical tongue twister condition to another condition with the same number of shifted and unshifted sounds outside of tongue twister contexts.<sup>7</sup> We take this to be an interesting direction for future work, but note that such a design would likely require even larger numbers of participants to be able to detect significant differences.

A second possibility is that the manipulations in our experiments, for whatever reasons, were not viewed as plausible causes for the type of atypical, shifted pronunciation we used. Specifically, the type of atypical pronunciation that participants heard in our experiments might not plausibly stem from speech errors because of, for example, faster speech rates, intoxication, or tongue twisters. Comparison with the acoustic properties of naturally occurring graded speech errors (as collected in, e.g., Alderete & Davies, 2018) would be required to

address this possibility. In the additional online data we report Experiment 5b, in which participants rated whether the stimuli from Experiments 1–5 involved production difficulty. Experiment 5b finds that tongue twisters were perceived as (somewhat) more likely to involve production difficulty only for phrases with shifted /f/, but not for phrases with shifted /s/. It is possible that we failed to find blocking of perceptual recalibration of /s/—the only shifted sound for which we found clear perceptual recalibration to begin with in the present experiments—because tongue twisters were not perceived as causing increased production difficulty for /s/.

A third possibility is that the *pattern* of atypical pronunciations we exposed participants to is perceived as unlikely to stem solely from incidental speech errors. In the present experiments, atypical pronunciations *always* occurred with the same sound (either always /s/ or always with /f/). This pattern might be objectively unlikely to occur if the atypical pronunciations are uncharacteristic of the talker, and just reflect incidental speech errors. For example, if the tongue twister contexts—that alternated between words with /s/- and words with /f/-onsets—are indeed the cause for the atypical pronunciations, why would speech errors always occur on just one of the two types of onsets (as is the case in our experiments)? Additionally, all of the atypical pronunciations in our experiments were about half-way shifted between the two phonemes /s/ and /f/. While phonetic blends *do* occur as the result of speech errors (e.g., Frisch & Wright, 2002; Goldstein et al., 2007; McMillan & Corley, 2010; Pouplier, 2007), it is unclear how likely it would be for four (Experiments 4 and 5) or even eight (Experiments 1 and 3) of such gradient speech errors to occur in sequence, in the absence of more categorical speech errors.

For example, in a large-scale study of natural speech, Alderete and Davies (2018) find that about 19% of speech errors are graded. This provides a lower-bound estimate how likely four such errors would be to occur in a row (lower bound, because gradient speech errors are particularly difficult to distinguish from the distribution of phonetic realizations that is expected even in the absence of any error). Simplifying somewhat, inference-based theories of perceptual recalibration (like the ideal adapter framework; Kleinschmidt & Jaeger, 2015) predict that the degree of change in the category boundary after an observation is a function of the observation's *improbability*. An observation can be probable either because it is probable as the result of an error or because it is probable under the distribution expected in the absence of an error. Future databases like those developed by Alderete and Davies (2018) should allow estimates of both of these components.

## Summary

The present results rule out naive causal inference accounts for blocking in perceptual recalibration. Either (a) perceptual recalibra-

<sup>6</sup> Independent of this possibility, the finding is compatible with a causal inference account, provided talkers still tend to be *more* likely to produce disfluencies before unfamiliar, complex references when they are distracted, even if distraction increases the overall frequency of disfluencies (see also Arnold et al., 2007, p. 928).

<sup>7</sup> It would further be important to mix shifted and unshifted sounds, because perceptual recalibration is known not to occur if all initial instances of a sound category produced by an unfamiliar talker are unshifted (Kraljic et al., 2008). Though perceptual recalibration experiments do not typically expose participants to mixtures of shifted and unshifted sounds, other paradigms have used mixtures of shifts and found boundary shifts closely resembling perceptual recalibration (e.g., Clayards et al., 2008; Kleinschmidt et al., 2015; Munson, 2011).



tion is unaffected by causal inferences and the result from earlier studies with visually presented incidental cues (pen in the mouth) are because of other mechanisms; (b) perceptual recalibration can be affected by causal inferences, but only for visually presented causes; or (c) perceptual recalibration is affected by causal inference regardless of the modality of the evidence, but the speech perception system is acutely attuned to what constitutes a plausible incidental cause for an observed deviation from expected pronunciations (and the present experiments failed to present sufficient plausible incidental causes). Variants of the paradigm we have developed here can be used in future work to distinguish between these three explanations.

## References

- Alderete, J., & Davies, M. (2018). Investigating perceptual biases, data reliability, and data discovery in a methodology for collecting speech errors from audio recordings. *Language and Speech, 62*, 281–317. <http://dx.doi.org/10.1177/0023830918765012>
- Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 914. <http://dx.doi.org/10.1037/0278-7393.33.5.914>
- Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: Disfluency and reference resolution. *Psychological Science, 15*, 578–582.
- Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America, 133*, EL174–EL180.
- Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science, 14*, 592–597. <http://dx.doi.org/10.1046/j.0956-7976.2003.psci.1470.x>
- Bicknell, K., Bushong, W., Tanenhaus, M. K., & Jaeger, T. (2019). *Listeners can maintain and rationally update uncertainty about prior words*. Manuscript submitted for publication.
- Bowers, J. S., Kazanina, N., & Andermane, N. (2016). Spoken word identification involves accessing position invariant phoneme representations. *Journal of Memory and Language, 87*, 71–83.
- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*, 707–729.
- Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*, 9–25.
- Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Maintaining information about speech input during accent adaptation. *PLoS ONE, 13*, e0199358.
- Bushong, W., & Jaeger, T. (2017). Maintenance of perceptual information in speech perception. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society (CogSci17)* (pp. 1129–1134). London, UK: Cognitive Science Society.
- Chin, S. B., & Pisoni, D. B. (1997). *Alcohol and speech*. Cambridge, MA: Academic Press.
- Choe, W. K., & Redford, M. A. (2012). The distribution of speech errors in multi-word prosodic units. *Laboratory Phonology, 3*, 5–26.
- Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116*, 3647–3658.
- Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception & Psychophysics, 70*, 604–618.
- Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*, 804–809.
- Cutler, A., & Henton, C. G. (2004). There's many a slip'twixt the cup and the lip. In H. Quene & V. v. Heuven (Eds.), *On speech and language: Studies for Sieb G. Nootboom* (pp. 37–45). Utrecht, the Netherlands: Netherlands Graduate School of Linguistics (LOT).
- Dix, S., Gardner, B., Lawrence, R., Morgan, C., Sullivan, A., & Kurumada, C. (2018). *Integration of top-down and bottom-up information in online interpretations of scalar adjectives*. Manuscript submitted for publication.
- 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*, 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*, EL307–EL313.
- Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology, 4*, 99–109.
- Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*, 224–238.
- Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America, 119*, 1950–1953.
- Ferber, R. (1991). Slip of the tongue or slip of the ear? On the perception and transcription of naturalistic slips of the tongue. *Journal of Psycholinguistic Research, 20*, 105–122.
- Frisch, S. A., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics, 30*, 139–162.
- Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language, 47*, 27–52.
- Garnham, A., Shillcock, R. C., Brown, G. D., Mill, A. I., & Cutler, A. (1981). Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics, 19*, 805–818.
- Goldrick, M., & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes, 21*, 649–683.
- Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition, 103*, 386–412.
- Grodner, D., & Sedivy, J. C. (2011). The effect of speaker-specific information on pragmatic inferences. In N. J. Pearlmuter & E. Gibson (Eds.), *The processing and acquisition of reference* (pp. 239–272). Cambridge: MIT Press.
- Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics, 48*, 865–892.
- Hay, J., Nolan, A., & Drager, K. (2006). From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review, 23*, 351–379.
- Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics, 34*, 458–484.
- Heigl, B. (2018). *[s] under the influence of alcohol*. Retrieved from [http://www.gmu.edu/org/lingclub/WP/texts/8\\_Heigl2.pdf](http://www.gmu.edu/org/lingclub/WP/texts/8_Heigl2.pdf)
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAS (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology, 61*, 23–62.
- Jaeger, T. F., Burchill, Z., & Bushong, W. (2019). *Strong evidence for expectation adaptation during language understanding, not a replication failure. A reply to Harrington Stack, James, and Watson (2018)*. Retrieved from <https://osf.io/4vxyp/>
- Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology, 15*, 281–319.
- Kleinschmidt, D., & Jaeger, T. F. (2011). A Bayesian belief updating model of phonetic recalibration and selective adaptation. In F. Keller & D. Reitter (Eds.), *Proceedings of the 2nd Workshop on Cognitive Mod-*

- eling and Computational Linguistics* (pp. 10–19). Portland, OR: Association for Computational Linguistics.
- Kleinschmidt, D. F., & Jaeger, T. F. (2012). A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci12)*. Sapporo, Japan.
- 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*, 148.
- Kleinschmidt, D. F., & Jaeger, T. F. (2016). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? *Psychonomic Bulletin & Review*, *23*, 678–691.
- Kleinschmidt, D. F., Raizada, R. D., & Jaeger, T. (2015). Supervised and unsupervised learning in phonetic adaptation. In D. Noelle, R. Dale, A. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci15)* (pp. 1129–1134). Austin, TX: Cognitive Science Society.
- Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*, 141–178.
- Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*, 262–268.
- Kraljic, T., & Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition*, *121*, 459–465.
- Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science*, *19*, 332–338.
- Kurumada, C., Brown, M., Bibyk, S., & Tanenhaus, M. K. (2018). *Probabilistic inferences and adaptation in pragmatic interpretation of contrastive prosody*. Manuscript submitted for publication.
- Lancia, L., & Winter, B. (2013). The interaction between competition, learning, and habituation dynamics in speech perception. *Laboratory Phonology*, *4*, 221–257.
- Levelt, W. J. (1993). *Speaking: From intention to articulation, Vol. 1*. Cambridge: MIT Press.
- Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition*, *174*, 55–70.
- MacKay, D. G. (1982). The problems of flexibility, fluency, and speed-accuracy trade-off in skilled behavior. *Psychological Review*, *89*, 483.
- McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech*, *58*, 502–521. <http://dx.doi.org/10.1177/0023830914565191>
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746.
- McMillan, C. T., & Corley, M. (2010). Cascading influences on the production of speech: Evidence from articulation. *Cognition*, *117*, 243–260.
- McMurray, B., Rhone, A., & Galle, M. (2012). Fricative Maker Pro [Computer software].
- Montero-Melis, G., Eisenbeiss, S., Narasimhan, B., Ibarretxe-Antuñano, I., Kita, S., Kopecka, A., . . . Bohnemeyer, J. (2017). Satellite-vs. verb-framing underpredicts nonverbal motion categorization: Insights from a large language sample and simulations. *Cognitive Semantics*, *3*, 36–61. <http://dx.doi.org/10.1163/23526416-00301002>
- Motley, M. T., & Baars, B. J. (1976). Laboratory induction of verbal slips: A new method for psycholinguistic research. *Communication Quarterly*, *24*, 28–34.
- Mowrey, R. A., & MacKay, I. R. (1990). Phonological primitives: Electromyographic speech error evidence. *The Journal of the Acoustical Society of America*, *88*, 1299–1312.
- Munson, C. M. (2011). *Perceptual learning in speech reveals pathways of processing* (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.
- Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*, 62–85.
- Nooteboom, S. G. (1980). Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V. A. Fromkin (ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 87–95). New York, NY: Academic Press.
- Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238.
- Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under- and over-informative pronominal adjective use. *Frontiers in Psychology*, *6*, 2035.
- Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition*, *77*, 97–132.
- Pouplier, M. (2007). Tongue kinematics during utterances elicited with the slip technique. *Language and Speech*, *50*, 311–341.
- 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*, 539–555.
- Rohde, H., & Kurumada, C. (2018). Alternatives and inferences in the communication of meaning. *Current Topics in Language*, *68*, 215.
- Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, *17*, 405–409.
- Samuel, A. G. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive Psychology*, *18*, 452–499.
- Samuel, A. G. (1989). Insights from a failure of selective adaptation: Syllable-initial and syllable-final consonants are different. *Perception & Psychophysics*, *45*, 485–493.
- Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology*, *32*, 97–127.
- 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.
- Sevold, C. A., & Dell, G. S. (1994). The sequential cuing effect in speech production. *Cognition*, *53*, 91–127.
- Shattuck-Hufnagel, S. (1983). Sublexical units and suprasegmental structure in speech production planning. In P. MacNeilage (Ed.), *The production of speech* (pp. 109–136). New York, NY: Springer.
- Shattuck-Hufnagel, S., & Klatt, D. H. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior*, *18*, 41–55.
- 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*, 3306–3316.
- Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition*, *96*, B13–B22.
- Tzeng, C. Y., Alexander, J. E., Sidaras, S. K., & Nygaard, L. C. (2016). The role of training structure in perceptual learning of accented speech. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 1793.
- Vroomen, J., & Baart, M. (2009). Recalibration of phonetic categories by lipread speech: Measuring aftereffects after a 24-hour delay. *Language and Speech*, *52*, 341–350.
- Vroomen, J., van Linden, S., De Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia*, *45*, 572–577.
- Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.
- Weatherholtz, K., & Jaeger, T. F. (2016). Speech perception and generalization across talkers and accents. *Oxford Research Encyclopedia of Linguistics*. Retrieved from <https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-95>

- Wijnen, F. (1992). Incidental word and sound errors in young speakers. *Journal of Memory and Language*, 31, 734–755.
- Wilshire, C. E. (1999). The “tongue twister” paradigm as a technique for studying phonological encoding. *Language and Speech*, 42, 57–82.
- Xie, X., Liu, L., & Jaeger, T. F. (2019). *Cross-talker generalization in foreign-accented speech perception*. Retrieved from [osf.io/brwx5](https://osf.io/brwx5)
- Xie, X., Theodore, R. M., & Myers, E. B. (2017). More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the

- internal structure of phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, 43, 206.
- Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America*, 143, 2013–2031.
- Zhang, X., & Samuel, A. G. (2014). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance*, 40, 200.

## Appendix

### Power Analyses

This section is best understood after reading Experiment 1. We outline our approach to the power analysis for Experiment 1. Because the effect sizes for the Label and Context effects we observe across experiments are rather constant, and since the number of subjects per condition (40 successful subjects) and the number and type of test items are held entirely constant across all experiments, the power estimates provided in the main text are representative for all experiments. The script for the power analysis is shared at <https://osf.io/ungba/>.

We conducted parametric generative power analysis (for examples of this approach, see Jaeger, Graff, Croft, & Pontillo, 2011; Montero-Melis et al., 2017). We used the same type of mixed logistics regression model used below to analyze the /s/ vs. /ʃ/ responses during the test block to generate 10,000 simulated data sets with hypothesized effects for Label, Context, and their interaction. Each of the 10,000 generated data sets was then analyzed in the same way as reported below. The goal of this was to determine whether we can detect (a) significant effects of perceptual recalibration (main effect of Label condition) and (b) significant blocking of perceptual recalibration (interaction of Label and Context conditions). Power for each of these effects was calculated as the percentage of times out of the 10,000 simulated data sets the underlying effect (present in the data generation process) was successfully detected.

As a conservative estimate of the effect of Label, we halved the Label effect observed in the first test block of Liu and Jaeger (2018;  $\beta = .56$  log-odds). As a conservative effect of the interaction between Label and Context condition, we used half the size of the Label effect ( $\beta = .28$ )—that is, our power analyses assess whether we would be able to detect a halving of the perceptual recalibration effect in the Tongue Twister condition, compared with the Non-Tongue Twister condition. As an additional conservative step, our power analyses pretend that we have only the data from the first block (i.e., seven instead of 35 test trials). Finally, we used a conservative (large) estimate for by-subject variance of the intercept, the only random effect in our analyses. Specifically, we set this variance to twice that observed in Experiment 1 ( $\sigma^2 = .9$ ).

These steps were taken to avoid over-optimism because of possibly inflated effect size estimates reported in previous work. We initially assumed both the intercept and the main effect of Context to have an effect of 0 log-odds. Here, we instead report power for a simulation

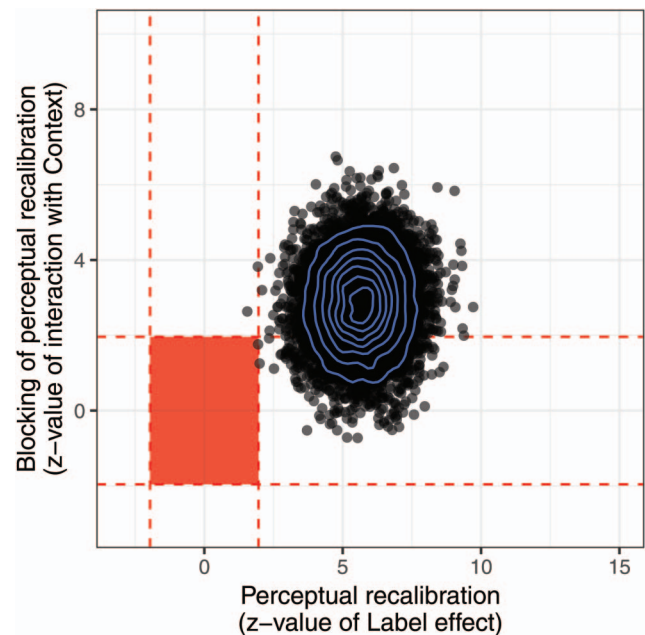


Figure A1. Distribution of  $z$ -values for the effect of Label and its interaction with Context across the 10,000 simulated data sets. Points outside of the red shaded rectangle indicate significant effect. Effects in the predicted direction have positive  $z$ -values. See the online article for the color version of this figure.

based on estimates for the intercept ( $\beta = -.25$  log-odds) and Context ( $\beta = -.07$ ) from Experiment 1, as this estimates the constraints of the present experiments more closely.

Figure A1 shows the distribution of  $z$ -values for the Label effect (perceptual recalibration) and its interaction with Context (the blocking of recalibration) across the 10,000 simulated data sets. As reported in the main text, power was very high for both effects (>95% for the Label effect and >81% for the interaction with Context).

Received December 29, 2018  
Revision received July 13, 2019  
Accepted July 17, 2019 ■