Transfer of statistical learning from speech perception to production generalizes to reading

Kyle D. Huffaker[1,3], Lori L. Holt[2], Nazbanou Nozari[1,3]

[1]Department of Psychological and Brain Sciences, Indiana University, [2]Department of Psychology, University of Texas at Austin, [3]Cognitive Science Program, Indiana University

Author Note

We have no conflicts of interest to disclose. This work was supported by National Science Foundation Grant BCS-2346989 to N. N. and L. L. H. and National Institutes of Health Award Number T32HD007475 to K. D. H. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Hamdaan Bhat for his significant help with accuracy checking across all experiments.

Correspondence concerning this article should be addressed to Kyle D. Huffaker, Department of Psychological and Brain Sciences, Indiana University, 1101 E 10th St., Bloomington, IN 47405. Email: kydahuff@iu.edu.

Transfer of statistical learning from speech perception to production generalizes to reading

Abstract

Past research has shown that short-term exposure to speech carrying certain acoustic statistics transfers robustly to speech production. However, all studies reporting such transfer have used auditory repetition tasks. Therefore, it is unclear whether perception-production transfer in the acoustic-phonetic domain extends to tasks without an auditory model to probe production. We answer this question in two experiments. Experiment 1 shows that people read aloud the words BEER and PEER differently after exposure to auditory samples of "beer" and "peer" drawn from a distribution of standard American English vs. a distribution of slightly accented speech. Experiments 2A and 2B replicate this finding and show generalization to reading a new word pair (BEACH/PEACH) and a new nonword pair (BEETH/PEETH). Collectively, these results demonstrate that the perception-production transfer in the acoustic-phonetic domain extends beyond auditory repetition tasks to production tasks without an explicit auditory model, and that this transfer generalizes to new syllables.

Keywords: speech perception, speech production, statistical learning, phonetic convergence, reading

Introduction

When interacting with speakers with an accent, we sometimes find ourselves speaking like them (Pardo et al. 2017). Despite a lifetime's experience speaking in our native accent, the speech patterns of an interlocutor can rapidly reshape our own (Murphy et al., 2024, 2025a, 2025b). For example, Murphy et al. (2024) showed that exposure to slightly accented "beer"/ "pier" utterances swiftly changed how listeners produced those words. In American English, /b/ is typically produced with a short voice onset time (VOT) and low fundamental frequency (F0), while /p/ is produced with a long VOT and high F0. A slight accent can be created by reversing these correlations, i.e., shifting /b/ signaled by short VOT to possess a high F0, and a /p/ with long VOT and a low F0 (Idemaru & Holt, 2011). Because VOT is a stronger acoustic cue than F0, listeners exposed to such reverse statistics still hear a /b/ as a /b/ and a /p/ as a /p/ but implicitly *downweight* F0 as an informative cue for the phoneme category. This is evident in the change to categorizing auditory stimuli when VOT is no longer an informative cue: When tested with VOT-ambiguous "beer"/ "pier" syllables after listening to typical American English accent, listeners to tend to report the stimuli with a low F0 as "beer" and those with a high F0 as "pier". However, after exposure to the accented speech, F0 no longer distinguishes between these two syllables, leading listeners to categorize F0-differentiated stimuli equally as /b/ and /p/ (Idemaru & Holt, 2011; Hodson et al., 2023). Murphy et al. (2024) extended this finding to production by asking listeners to repeat the test syllable. They found that, in line with the downweighting of F0 in perception, the difference in production F0s for "pier" and "beer" was reduced after exposure to accented speech. This finding was replicated in Murphy et al. (2025a, 2025b).

While the above studies show a robust transfer of changes from perception to production, all of them have used auditory repetition tasks that require the speaker to model an auditory probe.

A reasonable criticism of such studies is that the claim of transfer to "production" is overstated, as auditory repetition is not a pure production task. Importantly, it could be carried out by direct mapping of input to output phonology, bypassing lexical representations and the process of mapping those representations to their phonemes (Nozari et al., 2010; Nozari & Dell, 2013). It thus remains an open question: Is perception-production transfer also observed in production tasks that do not involve the auditory perception of a test stimulus? This paper answers this question by eliciting production through reading.

Examining transfer in the context of reading helps further isolate the cognitive component of perception-production transfer. In prior studies using the auditory repetition paradigm, both the exposure and the test stimuli carried some social information, such as the speaker's sex. A reading paradigm removes all social cues from the test stimulus, as it appears in text format. Such cues are known to modulate the convergence of one's speech to the interlocutor's speech. For example, convergence can vary as a function of the match or mismatch between the listener's and interlocutor's gender (Miller et al., 2010; Pardo et al., 2017). Similarly, a host of social factors, such as the speaker's perceived ethnicity, region, and social status can alter convergence (e.g., Bourhis & Giles, 1977; Giles et al., 1991; Pardo, 2006).

An additional advantage of reading over auditory repetition is the examination of generalization to syllables that were never heard in the study. Generalization is often assessed by exposing listeners to one pair (e.g., "bear", "pear") and testing them on a different pair (e.g., "beer", "pier"). By manipulating the overlap between the exposure and test pair, we can determine the representations involved in downweighting in perception and their transfer to production. Murphy et al. (2025b) tested two levels of generalization: phonemic and sub-phonemic. In the phonemic generalization task, the critical phonemes (i.e., /b/ and /p/) were shared between the exposure and

test pairs, but the vowel was different (e.g., bear/pear → beer/pier). Uncovering generalization at this level implies that the effects do not require overlap in full words, syllables or even a CV; rather, having the same phoneme is sufficient to reproduce the effect. In subphonemic generalization, the exposure and test pairs did not share the critical phonemes but shared the same critical dimensions (VOT and F0) and their correlations. For example, if the exposure pair contained /b/ and /p/ (e.g., "bear/pear"), the test pair contained /d/ and /t/ (e.g., "dear/tear"), where /d/ and /t/ have the same relationship to one another regarding VOT and F0 as /b/ and /p/. Uncovering generalization at this level implies that the effects do not depend on phonemic categories, but rather on the change to the representation of acoustic dimensions themselves.

In line with past studies (Idemaru & Holt, 2014, 2020), Murphy et al. (2025b) found generalization at the phonemic, but not subphonemic, level in perception. However, they did not find generalization to production at either level. One interpretation of this finding is that critical units of processing may be different in perception and production (Samuel, 2020). However, Murphy et al. (2025b) did not test CV-level generalization, where the exposure and test syllables share the CV- but not the coda C. Given coarticulation, production may show generalization across syllables, as long as the CV is shared between exposure and test pairs. Therefore, the current study targeted CV-level generalization in reading both words and nonwords.

The rationale for including both words and nonwords as test materials were two-fold. First, nonwords provide a robust test of CV-level generalization from learning across word stimuli in perception. For example, when the sort of perceptual downweighting studied by Murphy et al. (2025b) is elicited by word stimuli, it readily generalizes to nonsense syllables with the same phoneme onsets (Liu & Holt, 2015; Lehet & Holt, 2020; see also Kraljic & Samuel, 2005 for an example in another paradigm). Second, and more importantly, words and nonwords could tap into

different cognitive systems during reading. According to the dual-route cascaded (DRC) model of reading (Coltheart et al., 2001; Coltheart, 2006), words could be read using a lexical route (which may or may not reach semantic knowledge). In this route, visual features of letters activate letter representations, and those in turn, activate the word's orthographic representation in the lexicon. The orthographic representation then activates its corresponding phonological representation in the phonological lexicon, which then activates its corresponding phonemes. Nonwords, on the other hand, cannot be read this way because they do not have any stored representations in the orthographic or the phonological lexicon. DRC, thus, proposes a graphic-phoneme correspondence (GPC) route for reading nonwords. Rather than relying on stored representations in the lexicon, this route uses rules to convert letter strings directly into phoneme strings. Note that words could also be read through the GPC as long as they have regular spelling; however, reading through the lexical route is faster and more efficient.

The GPC route is close to the nonlexical route of auditory word repetition (Nozari et al., 2010), which directly maps input to output phonology, bypassing lexical representations. Here, too, lexical items can be produced using either route, although using the nonlexical route alone for repeating lexical items is uncommon (Nozari & Dell, 2013). Prior studies of perception-production transfer cannot disentangle the involvement of lexical vs. nonlexical route in transfer, because nonwords were not used in those studies. But understanding this issue is important in predicting the scope of the effect. It is possible that the perception-production effect depends on direct mapping of input to output phonology. Prior findings on word pairs should then be interpreted as them having been repeated primarily (although most likely not exclusively) through the nonlexical route, thus excluding the involvement of their stored lexical representation in production. If true, then any generalization effects are expected to be stronger for nonwords, which exclusively use

this route, compared to words. This possibility is theoretically interesting, but it decidedly limits the scope of transfer, as everyday speaking does not generally involve bypassing stored lexical representations. On the other hand, if transfer is present even when lexical representations are engaged, transfer should not show a large advantage for nonwords.

In summary, the current study had two goals: (a) to examine the basic perception-production transfer in reading, which does not have an auditory model (Experiment 1), and if uncovered, (b) to test generalization of such transfer to words and nonwords without participants having heard an auditory model for either (Experiments 2A, B).

Experiment 1

Method

Participants

A power analysis (PANGEA; Westfall, 2015) showed that, to detect a Condition x First Letter interaction with an effect size of 0.3 as the smallest effect size of interest, with a power of 0.8 and alpha of 0.05, a sample size of 33 was required. Accordingly, we recruited 44 participants using Prolific (www.prolific.com) to allow for attrition. Participants were adult English speakers in the U.S., aged 18-35 years with initial English exposure before age 2 years. All reported normal hearing. They were paid $10/hour for their time. Eleven participants were rejected due to background noise in recordings preventing F0 extraction or failure to comply with experimental instructions. Data from 33 participants (M_{Age} = 28.6, SD = 5; N_{Female} = 16) entered the analysis.

Stimuli

Materials consisted of auditory stimuli for the exposure phase and written words for the test phase. Exposure stimuli were audio recordings of the words *beer* and *peer* adopted from Murphy et al. (2025b). The specific tokens for *beer* and *peer* had a similar duration (400 ms) and F0 contour. Fundamental frequency (F0) at the onset of voicing was manipulated along with VOT to create a two-dimensional F0 x VOT acoustic space. F0 onset frequency was manipulated in 10-Hz steps in a 220-320 Hz range, and VOT was manipulated in 5-ms steps in a 5-40 ms range. Auditory stimuli were sampled from this two-dimension acoustic space to create two statistical regularity conditions: Canonical and Reverse. Figure 1A displays the distribution of stimuli in the two conditions. In the Canonical condition, which simulates the American English F0 x VOT relationship, /b/ was represented with a low F0 range of 220-240 Hz and a short VOT range of 5-15 ms and /p/ was represented with a high F0 range of 300-320 Hz and a long VOT range of 35-45 ms. In the Reverse condition, which simulated an accent, this mapping was flipped: stimuli with short VOT were placed in the high F0 range (heard as *beer*), and stimuli with long VOT (heard as *peer*) were placed in the low F0 range. We sampled multiple points in the F0 × VOT space allotted for each condition, allowing us to simulate variability in speech input while preserving the overall F0 × VOT correlation. Test stimuli were words BEER and PEER printed in uppercase black Calibri font (78 pt) on a white background.

Procedure

Online participants were recruited via Prolific and directed to the experiment hosted on Gorilla (www.gorilla.sc; Anwyl-Irvine et al., 2020). Participants completed consent and demographics forms before performing a headphone check (Milne et al., 2020) and a microphone check. Those who did not pass the headphone check after two attempts were rejected from the

experiment. Next, participants read instructions and watched a demo video of a sample trial with the orthographic test syllable BEEV, which was not presented in the experiment.

Each trial consisted of an auditory exposure phase and a word reading test phase. Figure 1B depicts the trial structure. In the auditory exposure phase, participants heard eight auditory stimuli randomized in order across trials, with four having a short-VOT and four having a long-VOT heard by most listeners as *beer* and *peer,* respectively. Immediately after, in the test phase, one target word appeared on screen and the participant read it aloud within a 2000 ms deadline before moving on to the next trial. Participants' recordings were saved as .weba files onto Gorilla's servers. The experiment consisted of 64 trials, divided into four equal blocks of 16 trials. Blocks alternated between Canonical and Reverse conditions, such that Blocks 1 and 3 used the Canonical stimuli, and Blocks 2 and 4 used the Reverse stimuli. Following Murphy et al. (2025a, 2025b), we chose to lead with a Canonical block to allow us to introduce the accent in the Reverse block as a shift from English expectations. Orthographic test stimuli were identical across blocks. Every eight trials, there was a 15 second break. Two lists were created with randomized order of trials. Each participant was randomly assigned to one of the two lists. The experiment took approximately 15 minutes to complete.

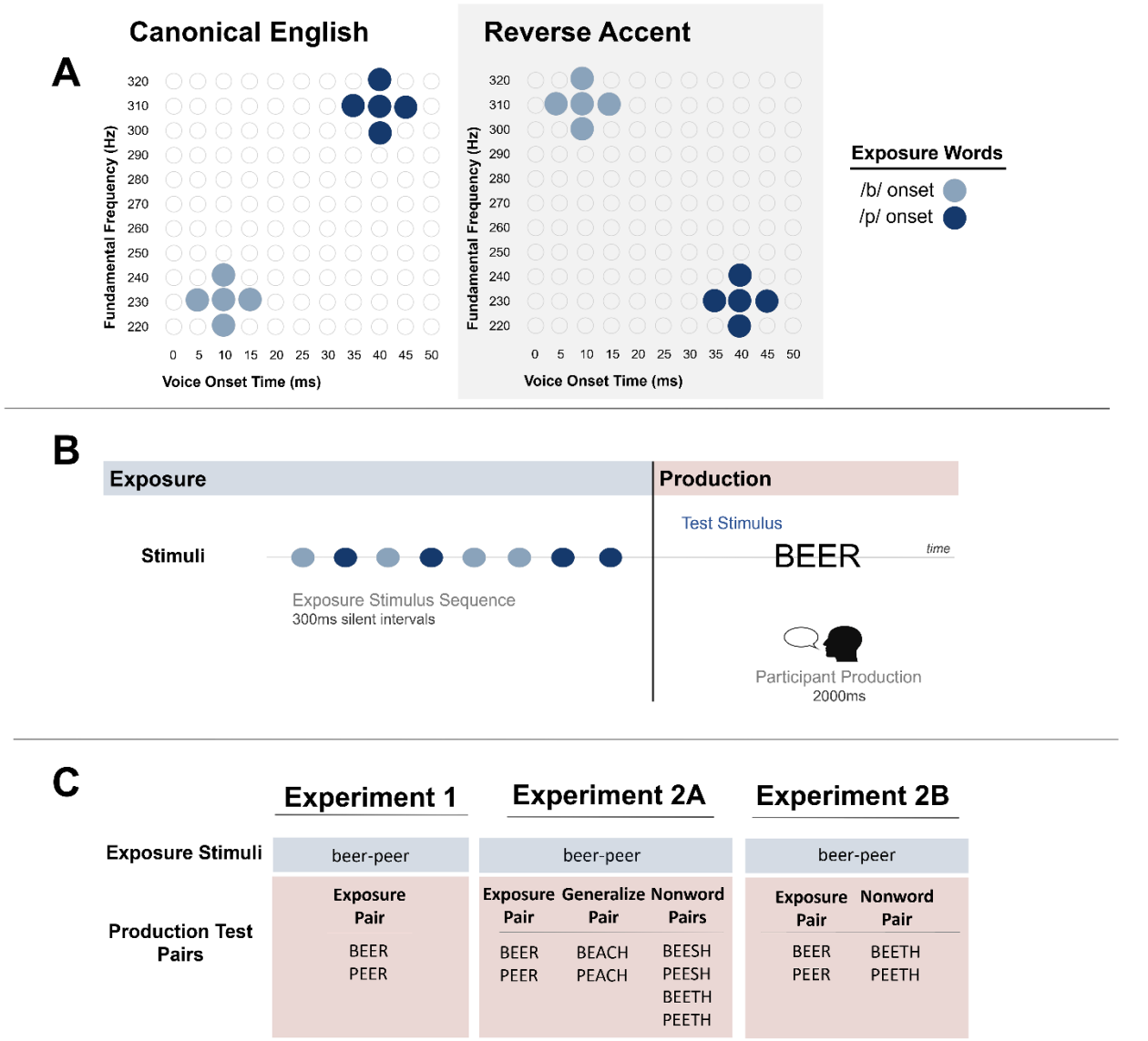

Figure 1. Outline of Experimental Design. **A.** Distribution of stimuli. Exposure stimuli are distributed across an acoustic space of voice onset time (VOT) and fundamental frequency (F0). Conditions are defined by a subsampling of this space that matches American English speech regularities (Canonical) or for which the VOT x F0 correlation is Reversed to create a subtle accent. **B.** Trial procedure. Participants heard eight words randomly sampled from the condition's stimulus distribution, with equal sampling from short- and long-VOT regions, and then read aloud a printed word **C.** Stimulus pairs in Experiments 1, 2A and 2B. In all experiments, participants heard *beer-peer* in exposure. Production Test Pairs included BEER-PEER, the lexical-generalization pair BEACH-PEACH, and the nonword generalization pairs BEESH-PEESH and BEETH-PEETH.

Analysis

Acoustic speech analysis was conducted using a custom pipeline described in Murphy et al. (2024). Trials with an incorrect or unclear response, disfluencies, murmuring or too much background noise were excluded from further analysis, as they do not provide reliable F0 measures. Acoustic speech analysis was conducted using a custom pipeline described in Murphy et al. (2024), using Praat (version 6.4.23; Boersma & Weenink, 2024). First, "To TextGrid (silences)..." identified and isolated word productions in the 2.5-second audio recordings. Next, "To Pitch (ac)" measured the F0 frequency of first 40 ms of voicing, where F0 differences between onset obstruent consonants are typically most pronounced (Hanson, 2009; Hombert et al., 1979; Lea, 1973; Xu & Xu, 2021). F0 measurements were made in 10 ms intervals, resulting in 5 samples per recording, which were averaged to generate one measurement per trial. Measurements that were ±3 standard deviations from the participant's average F0 across all productions were excluded from analysis. F0 was then normalized on a by-individual basis to mitigate the effects of between-subject variation in F0, such as sex (Titze, 1989). Thus, a *z* score of 0 represents the mean F0 for a participant across all their productions, and positive and negative *z* scores indicate higher or lower F0 relative to their mean, in standard deviation units. These *z* scores were the dependent measure of our analyses.

In general, we used generalized multi-level models with mixed effects whenever applicable. The random effect structure in Experiment 1 would have contained a random intercept of subjects, as well as random slopes for fixed effects over subjects. However, the model's estimation of random slopes was close to zero. Dropping random slopes would make the model similar to analysis of variance (ANOVA). Therefore, for this experiment, we used a 2 × 2 ANOVA with Condition (Canonical vs. Reverse) and First Letter (B vs. P) as fixed factors. ANOVA was

performed with the *afex* package (Version 1.4.1; Singmann et al., 2024) in R (Version 4.5.0; R Core Team, 2025). The data and results for all experiments are available on OSF.

Results

In total, 4.88% of responses were excluded, with similar exclusion rates across the Canonical and Reverse condition. Figure 2 displays the results (see Table A1 for descriptive statistics) and Table 1 presents the results of ANOVA.

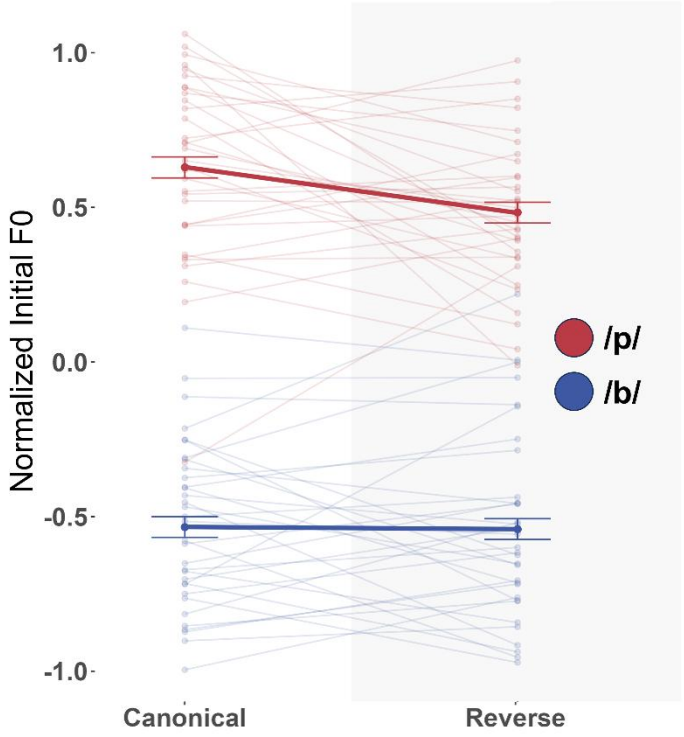

Figure 2. Experiment 1 results. Shows *z* score normalized initial F0 for reading-elicited production of *beer (blue) and peer (red)* across both conditions. The thick line indicates the sample mean; transparent lines show individual subjects' means. Error bars indicate ±1 standard error of the mean across subjects.

Table 1. 2 × 2 ANOVA of Experiment 1.

Measure	***MSE***	***F(1, 32)***	**Gη²**	***p***
Condition	0.07	2.88	.020	.099

First Letter	0.20	192.92	.796	< .001
Condition × First Letter	0.02	8.63	.016	.006

As expected, F0 was significantly higher in PEER than BEER, $F(1, 32) = 192.92$, $p < .001$. There was also a marginal main effect of Condition, $F(1, 32) = 2.88$, $p = .099$. Importantly for our purpose, there was a significant interaction between First Letter and Condition, $F(1, 32) = 8.63$, $p = .006$.

Discussion

Experiment 1 demonstrated that the statistics of the incoming auditory signal affect the production of written forms, without an auditory probe. Experiment 2A tested the generalization of this effect to lexical and nonlexical pairs.

Experiment 2A

Methods

Participants

A power analysis (PANGEA; Westfall, 2015) showed that a sample size of 39 was required to detect a three-way Condition × First Letter × Pair interaction with an effect size of 0.2 as the smallest effect size of interest, with a power of 0.85 and alpha of 0.05. Given the number of participants who had to be excluded due to background noise in Experiment 1, we recruited a larger sample of 65 participants for Experiment 2. Exclusion criteria and compensation were similar to Experiment 1. After exclusions, data from 50 participants ($M_{Age} = 26.87$, $SD = 6.14$; $N_{Female} = 28$) remained.

Stimuli

The Experiment 1 acoustic speech stimuli were used for the exposure phase. Test stimuli comprised four pairs. BEER-PEER was a lexical pair identical to the exposure stimuli (Word-Exposure), used to replicate the results of Experiment 1. BEACH-PEACH was a lexical pair different from the exposure stimuli, suitable for assessing generalization to other words that share a common onset with the exposure pair (Word-Generalization). BEESH-PEESH and BEETH-PEETH were nonlexical pairs, and thus obviously different from the exposure stimuli, suitable for assessing generalization to nonwords that share a common onset with the exposure pair (Nonword-Generalization). Two nonlexical pairs were used to balance the frequency of words and nonwords in the experiment, because different ratios of words and nonwords can change how participants read words (e.g., Hartsuiker et al., 2005). As such, each test pair consisted of 25% of the total trials. All words appeared on screen in uppercase 78 pt black Calibri font on a white background.

Procedure

The design was similar to Experiment 1. Experiment 2A consisted of 256 trials, with two blocks divided equally into 128 trials. There was a 15 second break every 8 trials, with a longer 60 second break every 32 trials. Block 1 was Canonical, and Block 2 was Reverse. Within each block, each of the 8 possible test words appeared 16 times. Three orders were created with a pseudorandomized trial order, with the following constraints: (1) words in the same test pair could not appear in subsequent trials, (2) no more than 3 words in a row could start with the same first letter, and (3) the lists were evenly divided into four quarters and each test word appeared a total of 8 times in each quarter. Participants were randomly assigned to one of the three orders. Participants completed the same trial procedure as before. The experiment took approximately 50 minutes to complete.

Analysis

We used the same F0 extraction and normalization procedure as in Experiment 1. Data were analyzed using linear mixed-effect models implemented in the *lme4* package with the lmer function (Bates et al., 2015) in R (Version 4.5.0; R Core Team, 2025). The dependent variable was normalized F0. Fixed effects included Condition (Canonical vs. Reverse) and First Letter (B vs. P). To measure generalization, we included two additional factors: Pair (BEACH-PEACH vs. BEER-PEER) to evaluate transfer to novel lexical items, and Lexicality (Word vs. Nonword) to test generalization from words to nonwords. All fixed effects were sum-coded (-1, 1). *P* values were calculated using Satterthwaite approximations via the *lmerTest* package (Version 3.1.3; Kuznetsova et al., 2016). We used the largest random effect structure that could be tolerated by each model.

Three sets of analyses were conducted on the data. Set 1 was a general analysis to test for the downweighting of F0 across all four pairs. The model had normalized initial Production F0 as its dependent variable. The fixed-effect structure included Condition (Canonical vs. Reverse) and First Letter (B vs. P), as well as the 2-way interaction between them. The random-effect structure included the random intercept of subjects, as well as the random slopes of Condition and First Letter over subjects.

Set 2 examined lexical generalization. Correspondingly, it used the subset of data with lexical pairs (BEER-PEER and BEACH-PEACH). The model had normalized initial Production F0 as its dependent variable. The fixed-effect structure included Condition (Canonical vs. Reverse), First Letter (B vs. P), Pair (BEER-PEER vs. BEACH-PEACH), as well as all the 2- and 3-way interactions between them. The random-effect structure included the random intercept of subjects, as well as the random slopes of Condition, First Letter, and Pair over subjects. We followed up on this analysis by reporting the results of two post-hoc tests conducted individually

on BEER-PEER and BEACH-PEACH subsets to investigate F0 downweighting in each pair individually.

Finally, Set 3 examined generalization to nonlexical items. Given the results of Set 2, which found very similar patterns between BEER-PEER and BEACH-PEACH, the final set included all four pairs to have the most balanced analysis. This model had normalized initial production F0 as its dependent variable. The fixed-effect structure included Condition (Canonical vs. Reverse), First Letter (B vs. P), Lexicality (Word vs. Nonword), as well as the 2- and 3-way interactions between them. The random-effect structure included the random intercept of subjects, as well as the random slopes of Condition, First Letter, and Lexicality over subjects. We followed this analysis with a post-hoc test on BEESH-PEESH and BEETH-PEETH combined to investigate downweighting in nonlexical pairs.

Results

In total, 13.63% of trials were excluded due to erroneous, unclear, or noisy productions, with similar exclusion rates across the Canonical and Reverse conditions. Figure 3 shows the results of Experiment 2A (see Table A1 for descriptive statistics).

Set 1. Table 2 shows the results of the general analysis. Across all trials, F0 was significantly higher in /p/ utterances than /b/ utterances ($\beta = -0.4864$, $t = -15.067$, $p < .001$). The main effect of Condition was not significant ($\beta = 0.0136$, $t = 0.752$, $p = .455$). Importantly for our purpose, there was a significant interaction between First Letter and Condition across all trials ($\beta = -.02449$, $t = -3.542$, $p < .001$).

Table 2. Results of the Set 1 analysis for Experiment 2A.

Predictor	**β**	***SE***	***t***	***p***
Intercept	0.003	0.010	0.296	.768
Condition	0.014	0.018	0.752	.455
First Letter	-0.486	0.032	-15.067	< .001
Condition × First Letter	-0.025	0.007	-3.542	< .001

Set 2. Table 3 shows the results of the lexical-generalization analysis. As expected, F0 was significantly higher for /p/ than /b/ (β = -0.499, t = -15.837, p < .001). There was also a main effect of Pair, with F0 overall higher in BEACH-PEACH than in BEER-PEER (β = 0.109, t = 5.163, p < .001). Importantly for our purpose, there was a significant interaction between Condition and First Letter, marking F0 downweighting (β = -0.034, t = -3.472, p < .001). However, downweighting did not interact with Pair (β = -0.004, t = -0.444, p = .657). Follow-up tests showed a significant main effect of First Letter for both BEER-PEER and BEACH-PEACH pairs (BEER-PEER: (β = -0.494, t = -14.802, p < .001); BEACH-PEACH: (β = -0.519, t = -13.945, p < .001). Importantly, they showed a significant Condition × First letter interaction in BEER-PEER (β = –0.030, t = –2.430, p = .015) replicating the results of Experiment 1 and a similar interaction in BEACH-PEACH (β = –0.037, t = –2.380, p = .017), showing generalization to a new lexical pair.

Set 3. Table 4 shows the results of the nonlexical-generalization analysis. Because of similar effect sizes demonstrated by the post-hoc analyses in Set 2, we chose to include BEACH-PEACH trials in this model to maintain symmetry between factors. As expected, F0 was significantly higher for /p/ than /b/ (β = -0.487, t = -15.062, p < .001). There was also a main effect of Lexicality, meaning that F0 was overall higher in BEESH-PEESH and BEETH-PEETH than in BEER-PEER and BEACH-PEACH (β = 0.038, t = 3.117, p = .003). There was a significant interaction between Condition and First Letter (β = -0.025, t = -3.610, p < .001), replicating the

transfer effect across the full dataset. Overall downweighting did not interact with Lexicality (β = 0.009, t = 1.275, p = .202), suggesting that the change in participant F0 from Canonical to Reverse conditions was similar for both Words and Nonwords. A follow-up test on the subset of all nonlexical trials showed a significant main effect of First Letter (β = -0.470, t = -13.12, p < .001). However, there was a marginal Condition × First Letter interaction (β = -0.016, t = -1.68, p = 0.093). While the overall model suggests that transfer may extend to Nonwords, the marginal subset result leaves this conclusion uncertain.

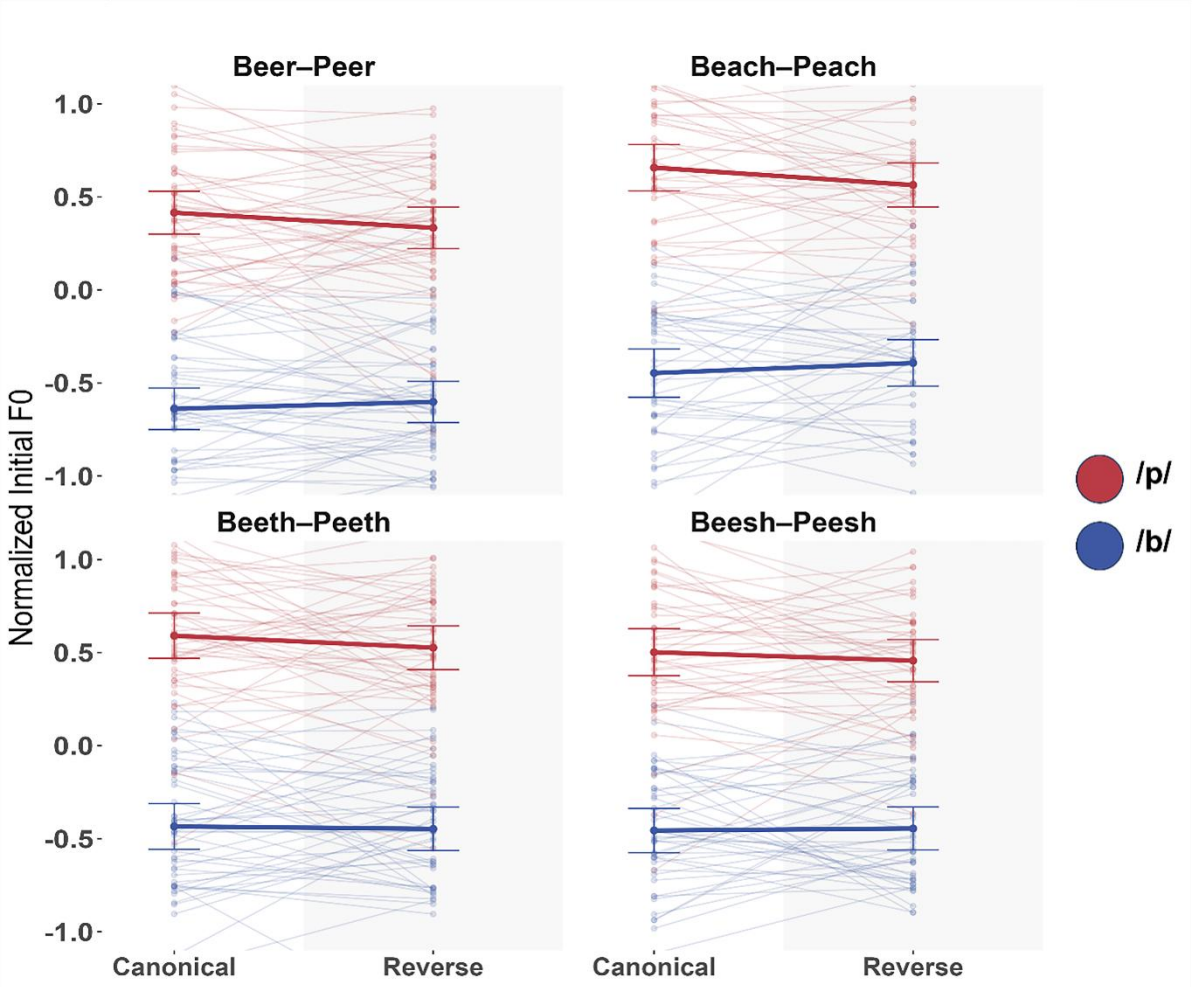

Figure 3. Experiment 2A results. Shows *z* score normalized initial F0 for participant production of /b/ *(blue) and* /p/ *(red)* across both conditions in all four test pairs. The thick line indicates the sample mean; transparent lines show individual subjects' means. Error bars indicate ±1 standard error of the mean across subjects.

Table 3. Results of the Set 2 (lexical generalization) analysis for Experiment 2A.

Predictor	**β**	***SE***	***t***	***p***
Intercept	-0.014	0.015	-0.956	.344
Condition	0.013	0.018	0.726	.471
First Letter	-0.499	0.031	-15.837	< .001
Pair	0.109	0.021	5.163	< .001
Condition × First Letter	-0.034	0.010	-3.472	< .001
Condition × Pair	0.001	0.010	0.048	.961

First Letter × Pair	-0.005	0.010	-0.527	.598
Condition × First Letter × Pair	-0.004	0.010	-0.444	.657

Table 4. Results of the Set 3 (nonlexical generalization) analysis for Experiment 2A.

Predictor	**β**	***SE***	***t***	***p***
Intercept	-0.000	0.011	-0.031	.975
Condition	0.014	0.018	0.759	.451
First Letter	-0.487	0.032	-15.062	< .001
Lexicality	0.038	0.012	3.117	.003
Condition × First Letter	-0.025	0.007	-3.610	< .001
Condition × Lexicality	0.001	0.007	-0.083	.933
First Letter × Lexicality	0.012	0.007	1.675	.094
Condition × First Letter × Lexicality	0.009	0.007	1.275	.202

Discussion

In summary, Experiment 2A replicated the results of Experiment 1, and showed clear generalization to a new lexical pair with a magnitude comparable to that of the original pair. However, while there was no significant interaction between downweighting and lexicality, downweighting in the combined set of nonlexical pairs was only marginal, preventing us from drawing clear conclusions about the scope of generalization. There are two possibilities for the marginal effect of nonlexical generalization. First, generalization may be weak or unstable for nonlexical pairs. Second, generalization may be solid for such pairs, but the data may have been too noisy to detect it. The latter is a distinct possibility, as the inclusion of 16 syllables in a much longer experiment led to many more exclusions due to errors and unclear productions, compared to Experiment 1. One pair, BEESH-PEESH, was specifically error-prone, generating 39% of errors made across all four pairs. We thus designed Experiment 2B as a shorter version of Experiment

2A, focusing on contrasting BEER-PEER with the cleaner nonword pair BEETH-PEETH. If there is generalization to nonlexical pairs, we expect Experiment 2B to show such generalization clearly.

Experiment 2B

Methods

Participants

Given the similarity in the goals of Experiments 2A and 2B, the same sample size calculation was employed, requiring 39 subjects. We recruited 53 subjects, and used 43 after excluding those who did not follow task instructions or had noisy backgrounds (M_{Age} = 30.11, SD = 4.18; N_{Female} = 22).

Stimuli

The materials were the same as Experiment 2A but only included BEER-PEER and the nonword BEETH-PEETH.

Procedure

The procedure matched Experiments 1 and 2A. Experiment 2B consisted of 128 trials, with two equal blocks of 64 trials each. Block 1 was the Canonical condition, and Block 2 was the Reverse condition. In each block, each of the four possible test words appeared 16 times. For the experiment, two test orders were created. Each order was split into equal quarters, and each test word was used an equal number of times per quarter. Participants were randomly assigned to an order. The experiment took approximately 25 minutes to complete.

Analysis

The same analysis as Set 3 of Experiment 2A was applied here. The model had normalized F0 as its dependent variable. The fixed-effect structure included Condition (Canonical vs. Reverse), First Letter (B vs. P), and Lexicality (Word vs. Nonword) as well as the 2- and 3-way interactions between them. The random-effect structure included the random intercept of subjects, as well as the random slopes of Condition and First Letter over subjects. We followed up on this analysis by conducting a post-hoc test on both BEER-PEER and BEETH-PEETH.

Results

In total, 4.09% of trials were excluded. In the Canonical condition, 3.99% of trials were excluded. In the Reverse condition, 4.17% of trials were excluded. Figure 4 shows the pattern of results in Experiment 2B and Table 5 summarizes the results of the analysis. As before, we found significantly higher F0s for /p/ vs. /b/ ($\beta = -0.412$, $t = -11.548$, $p < .001$). F0 was also significantly higher for BEETH-PEETH than BEER-PEER ($\beta = 0.159$, $t = 14.569$, $p < .001$). Importantly, we saw a significant interaction between Condition and First Letter, denoting downweighting ($\beta = -0.035$, $t = -3.224$, $p = .0012$), which did not interact with Lexicality ($\beta = -0.003$, $t = -0.267$, $p = .789$).

Table 5. Lexicality model results for Experiment 2B.

Predictors	**β**	***SE***	***t***	***p***
Intercept	0.002	0.011	0.181	.856
Condition	-0.001	0.031	-0.046	.963
First Letter	-0.412	0.036	-11.548	< .001
Lexicality	0.159	0.011	14.569	< .001
Condition × First Letter	-0.035	0.011	-3.224	.001
Condition × Lexicality	0.016	0.011	1.485	.138
First Letter × Lexicality	0.020	0.011	1.855	.064
Condition × First Letter × Lexicality	-0.003	0.011	-0.267	.789

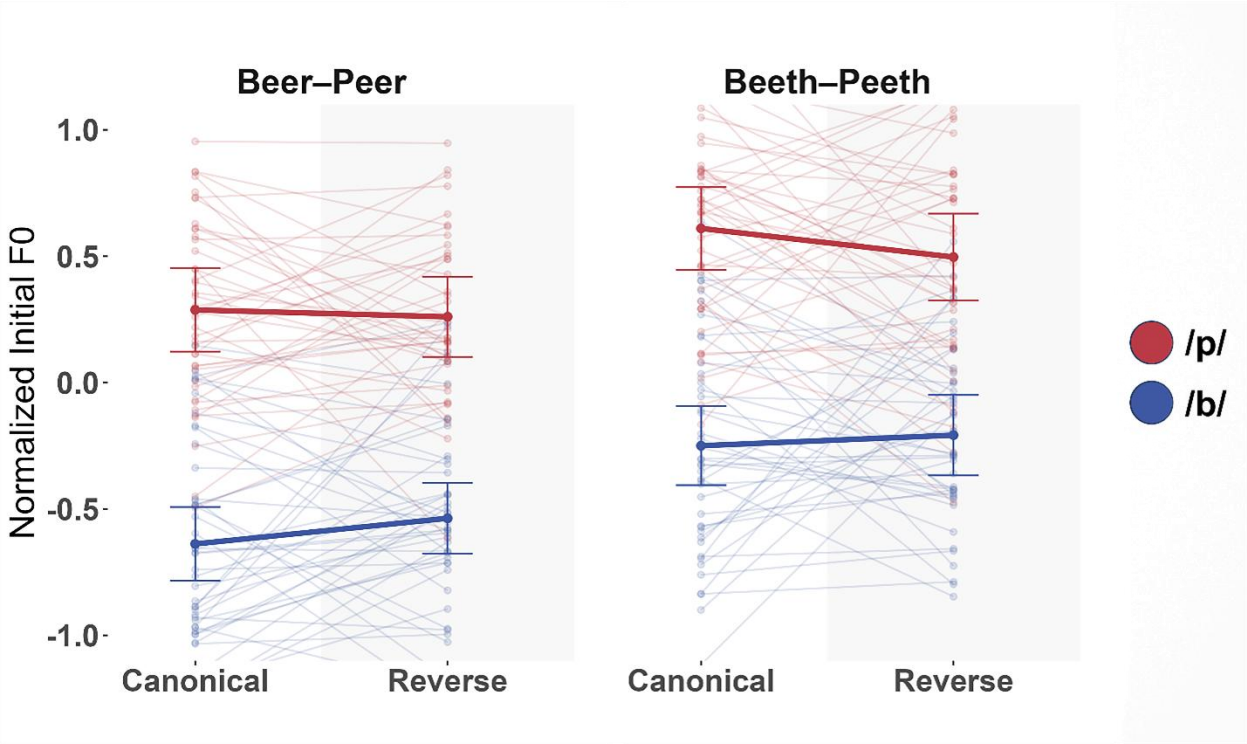

Figure 4. Experiment 2B results. score normalized initial F0 for participant production of /b/ *(blue) and* /p/ *(red)* across both conditions. The thick line indicates the sample mean; transparent lines show individual subjects' means. Error bars indicate ±1 standard error of the mean across subjects.

Post-hoc tests showed higher F0 for /p/ vs. /b/ in both BEER-PEER and BEETH-PEETH (BEER-PEER: $\beta = -0.432$, $t = -12.084$, $p < .001$; BEETH-PEETH: $\beta = -0.394$, $t = -9.716$, $p < .001$). Critically, they also showed a significant interaction between Condition and First Letter for BEER-PEER ($\beta = -0.031$, $t = -2.093$, $p = .0365$), providing a third replication of F0 downweighting in reading, as well as BEETH-PEETH ($\beta = -0.038$, $t = -2.447$, $p = .0145$), showing clear generalization to a nonlexical pair.[1]

[1] See Appendix B for a set of cross-experiment analyses testing the influence of pure lexical vs. mixed lexical-nonlexical contexts on word reading.

General Discussion

Experiment 1's results show that the perception-production transfer previously reported in auditory repetition tasks (Murphy et al., 2024, 2025a, 2025b) extends to reading. We replicated this effect two more times in Experiments 2A and 2B, leaving no room for doubt that an auditory prompt was not necessary for changes to the production system after exposure to new statistics in the perception system. We further showed generalization to new pairs that had never been auditorily modeled in the study. Based on past studies that had limited generalization in perception to shared phonemes (Idemaru & Holt, 2014, 2020; Murphy et al., 2025b), we calibrated our generalization experiments to this level and examined generalization to both words and nonwords.

Experiment 2A found generalization to a new word pair. Moreover, the magnitude of the transfer effect in the new pair was comparable to that of the exposure pair, suggesting that a shared CV, rather than the full syllable (CVC), is sufficient to produce generalization. In contrast, only a marginal generalization effect was obtained for the nonlexical pair in this experiment. As this result did not match either of the two theoretical possibilities discussed in the Introduction, we surmised that the marginal effect may have stemmed from low signal-to-noise ratio. Indeed, one of the two nonlexical pairs (BEESH-PEESH) was often read with errors and disfluencies, leading to the exclusion of a large number of trials. Given that the initial F0 is already quite variable in production (Murphy et al., 2024), exclusion of too many trials can introduce substantial noise into the analyses. Experiment 2B examined this possibility by focusing on the cleaner nonlexical pair (BEETH-PEETH). This experiment replicated the generalization observed in Experiment 2A, but this time with the nonlexical pair. Also similar to Experiment 2A, the size of the transfer effect was comparable between the exposure and the nonword pair, identifying the phoneme (and not the full CVC syllable) as the critical representation involved in transfer.

To summarize, we found perception-production transfer in reading that generalized readily to both new lexical and nonlexical stimuli that shared the critical phoneme with the exposure pair. Contrary to the prediction of a GPC-driven process, generalization was not weaker for words than nonwords. This finding implies that the change to the production system affects post-lexical (most likely articulatory-phonetic) representations, regardless of how they are accessed (lexically or sublexically). This is important for extending the scope of transfer from laboratory studies to everyday conversations. If transfer were limited to the sublexical route, it would significantly limit the possibility of observing transfer in more naturalistic contexts, where articulation is driven primarily by lexical access.

The uncovering and replication of generalization in the current study is different from the null generalization effect in production reported in Murphy et al. (2025b). The most likely reason for this difference is that the unit of generalization in production is the CV (current study) rather than the phoneme (Murphy et al., 2025b). This interpretation is supported by studies emphasizing the unique role of syllables in production during phonetic encoding (e.g., Cholin et al., 2006; Laganaro & Alario, 2006). Although we cannot, based on current results, rule out a role for task in explaining the differences in generalization between the current study and Murphy et al. (2025b), there is no clear theoretical reason why task should affect generalization, especially when the basic effect was robustly replicated across tasks. Nevertheless, we propose that future studies compare generalization at similar levels across tasks within the same individuals to establish the robustness and scope of generalization.

Declarations

Funding

This work was supported by the National Science Foundation Grant BCS-2346989 and the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number T32HD007475. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

The authors have no relevant financial or non-financial interests to disclose.

Ethics Approval

The study involving human data was approved by the Institutional Review Board of Indiana University Bloomington (Approval No. 19756). Informed consent for the use of their data was obtained from all participants.

Consent to Participate

Informed consent was obtained from all individual participants included in the study.

Consent for Publication

Not applicable.

Availability of Data and Materials

The data and tables used in the study are available on OSF:

https://osf.io/kusr4/?view_only=47488caa1de14739a87afec68a380d5b

Code Availability

The scripts used for statistical analyses in the study are available on OSF:

https://osf.io/kusr4/?view_only=47488caa1de14739a87afec68a380d5b

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Boersma, P., & Weenink, D. (2024). Praat: Doing phonetics by computer (version 6.4.23) [Computer software]. http://www.praat.org

Bourhis, R. Y., & Giles, H. (1977). The language of intergroup distinctiveness. *Language, Ethnicity and Intergroup Relations*, 13, 119.

Cholin, J., Levelt, W. J., & Schiller, N. O. (2006). Effects of syllable frequency in speech production. *Cognition*, *99*(2), 205-235.

Coltheart, M. (2006). Dual route and connectionist models of reading: An overview. *London Review of Education*, *4*(1), Article 1. https://doi.org/10.1080/13603110600574322

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256. https://doi.org/10.1037/0033-295X.108.1.204

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of Accommodation: Developments in Applied Sociolinguistics* (pp. 1–68). Cambridge University Press. https://doi.org/10.1017/CBO9780511663673.001

Hanson, H. M. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. The Journal of the Acoustical Society of America, 125(1), 425–441.

Hartsuiker, R. J., Corley, M., & Martensen, H. (2005). The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related beply to Baars et al. (1975). *Journal of Memory and Language*, *52*(1), 58–70. https://doi.org/10.1016/j.jml.2004.07.006

Hodson, A. J., Shinn-Cunningham, B. G., & Holt, L. L. (2023). Statistical learning across passive listening adjusts perceptual weights of speech input dimensions. *Cognition*, *238*, 105473. https://doi.org/10.1016/j.cognition.2023.105473

Hombert, J. M., Ohala, J. J., & Ewan, W. G. (1979). Phonetic explanations for the development of tones. Language, 55(1), 37–58.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939–1956. https://doi.org/10.1037/a0025641

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1009–1021. https://doi.org/10.1037/a0035269

Idemaru, K., & Holt, L. L. (2020). Generalization of dimension-based statistical learning. *Attention, Perception, & Psychophysics*, *82*(4), 1744–1762. https://doi.org/10.3758/s13414-019-01956-5

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178. https://doi.org/10.1016/j.cogpsych.2005.05.001

Laganaro, M., & Alario, F. X. (2006). On the locus of the syllable frequency effect in speech production. *Journal of Memory and Language*, *55*(2), 178-196.

Lea, W. A. (1973). Segmental and suprasegmental influences on fundamental frequency contours. Consonant Types and Tone, 1, 15–70.

Lehet, M., & Holt, L. L. (2020). Nevertheless, it persists: Dimension-based statistical learning and normalization of speech impact different levels of perceptual processing. *Cognition*, *202*, 104328. https://doi.org/10.1016/j.cognition.2020.104328

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1783–1798. https://doi.org/10.1037/xhp0000092

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2010). Alignment to visual speech information. *Attention, Perception, & Psychophysics*, *72*(6), 1614–1625. https://doi.org/10.3758/APP.72.6.1614

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, *53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Murphy, T. K., Nozari, N., & Holt, L. L. (2024). Transfer of statistical learning from passive speech perception to speech production. *Psychonomic Bulletin & Review*, *31*(3), 1193–1205. https://doi.org/10.3758/s13423-023-02399-8\

Murphy, T., Holt, L. L., & Nozari, N. (2025a). *Exposure to an Accent Transfers to Speech Production in a Single Shot* (SSRN Scholarly Paper No. 5196109). Social Science Research Network. https://doi.org/10.2139/ssrn.5196109

Murphy, T., Nozari, N., & Holt, L. L. (2025b). Bears don't always mess with beers: Limits on generalization of statistical learning in speech. Psychonomic Bulletin & Review. https://doi.org/10.3758/s13423-025-02690-w

Nozari, N., Kittredge, A. K., Dell, G. S., & Schwartz, M. F. (2010). Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of Memory and Language*, *63*(4), 541–559. https://doi.org/10.1016/j.jml.2010.08.001

Nozari, N., & Dell, G. S. (2013). How damaged brains repeat words: A computational approach. *Brain and Language*, *126*(3), 327–337. https://doi.org/10.1016/j.bandl.2013.07.005

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393. https://doi.org/10.1121/1.2178720

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*(2), 637–659. https://doi.org/10.3758/s13414-016-1226-0

R Core Team. (2025). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, *111*, 104070. https://doi.org/10.1016/j.jml.2019.104070

Singmann, H., Bolker, B., Westfall, J., Aust F., & Ben-Shachar M. (2024). *afex: Analysis of Factorial Experiments*. R package version 1.4-1, https://CRAN.R-project.org/package=afex

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, *85*(4), 1699–1707. https://doi.org/10.1121/1.397959

Westfall, J. (2015). PANGEA: Power analysis for general ANOVA designs. Unpublished manuscript. Available at http://jakewestfall.org/publications/pangea.pdf, 4

Xu, Y., & Xu, A. (2021). Consonantal F0 perturbation in American English involves multiple mechanisms. *The Journal of the Acoustical Society of America*, *149*(4), 2877–2895. https://doi.org/10.1121/10.0004239

Appendix A: Descriptive statistics across experiments

Table A1. Mean normalized F0 (in *z* scores) for Canonical and Reverse conditions in each experiment, along with the within-subject shift in F0, computed as the difference between Canonical and Reverse conditions. Standard errors (in parentheses) are computed over subject-level means.

Experiment	***Participants***	***Canonical (SE)***	***Reverse (SE)***	***Shift (SE)***
Exp 1	33	1.155 (0.083)	1.018 (0.079)	0.137 (0.046)
Exp 2A	50	1.021 (0.071)	0.915 (0.063)	0.106 (0.036)
Exp 2B	43	0.895 (0.080)	0.756 (0.071)	0.139 (0.051)

Appendix B: Cross-experiment analyses

As alluded to earlier, different ratios of words and nonwords could change how participants read words (e.g., Hartsuiker et al., 2005). In Experiment 1, participants were only exposed to words, whereas in Experiments 2A and 2B, they were exposed to a balanced mixture of words and nonwords. It is thus possible that participants have treated the word pairs as nonwords in the latter two experiments. This, in turn, could change the size of transfer for word pairs. To investigate this issue, we have selected the common subset of all three experiments, i.e., "BEER-PEER" pairs, and statistically compared the size of transfer for this pair in Experiment 1 vs. Experiment 2A (Table B1) and Experiment 2B (Table B2). Models were similar in structure to those described in detail in the main text, with the exception of the added variable of "Context" (Pure lexical vs. Mixed lexical-nonlexical) to the fixed structure and its interactions with the other two fixed effects, Condition and First Letter. The models' random effect structure included the random intercept of subjects, as well as the random slopes of Condition and First Letter.

Table B1. BEER-PEER downweighting compared between Experiment 1 and Experiment 2A.

Predictor	**β**	***SE***	***t***	***p***
Intercept	-0.055	0.017	-3.267	.002
Condition	0.027	0.015	1.755	.083
First Letter	-0.519	0.026	-20.049	< .001
Context	0.068	0.017	4.045	< .001
Condition × First Letter	-0.032	0.010	-3.219	.001
Condition × Context	0.015	0.015	0.971	.334
First Letter × Context	-0.025	0.026	-0.968	.336
Condition × First Letter × Context	-0.002	0.010	-0.201	.840

Table B2. BEER-PEER downweighting compared between Experiment 1 and Experiment 2B.

Predictor	**β**	***SE***	***t***	***p***
Intercept	-0.072	0.015	-4.961	< .001

Condition	0.012	0.020	0.613	.542
First Letter	-0.488	0.027	-18.346	< .001
Context	0.085	0.015	5.844	< .001
Condition × First Letter	-0.033	0.011	-2.963	.003
Condition × Context	0.029	0.020	1.482	.142
First Letter × Context	-0.056	0.027	-2.111	.038
Condition × First Letter × Context	-0.001	0.011	-0.124	.902

In both analyses, there was a main effect of Context, with average F0 higher in Experiment 1 compared to both Experiments 2A and 2B, most likely reflecting the differences in the samples of these experiments. However, the effect of interest to us is the three-way interaction between Condition x First Letter and Context. This interaction was nowhere near significant in either analysis. To make sure that this null statistical result was not due to low statistical power, we conducted a third analysis, aggregating the data from Experiments 2A and 2B (this aggregation was sanctioned by the similar pattern observed in the two analyses reported above). The results are reported in Table B3.

Table B3. BEER-PEER downweighting compared between Experiment 1 and the aggregate BEER-PEER responses across Experiments 2A and 2B.

Predictor	**β**	***SE***	***t***	***p***
Intercept	-0.063	0.015	-4.098	< .001
Condition	0.020	0.016	1.249	.214
First Letter	-0.505	0.024	-21.404	< .001
Context	0.076	0.015	4.944	< .001
Condition × First Letter	-0.033	0.009	-3.472	< .001
Condition × Context	0.021	0.016	1.335	.184
First Letter × Context	-0.039	0.023	-1.671	.097
Condition × First Letter × Context	-0.002	0.009	-0.185	.853

As can be seen in the table, the three-way interaction remained non-significant. We can, thus, conclude that the lexicality of context did not significantly modulate the size of the transfer effect for lexical items.