Bears Don't Always Mess with Beers: Limits on Generalization of Statistical Learning in Speech

Timothy K. Murphy[1,2], Nazbanou Nozari[3,4], and Lori L. Holt[5]

1. Waisman Center, University of Wisconsin-Madison, Madison, WI, USA
2. Department of Surgery, University of Wisconsin-Madison, Madison, WI, USA
3. Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, USA
4. Cognitive Science Program: Indiana University, Bloomington, IN, USA
5. Department of Psychology, The University of Texas at Austin, Austin, TX, USA

Corresponding Author.
Timothy K. Murphy
Waisman Center, 1500 Highland Ave, Madison, WI 53705
Email: tmurphy37@wisc.edu
(484) 888-1868

Abstract

Perception changes rapidly and implicitly as a function of passive exposure to speech that samples different acoustic distributions. Past research has shown that this statistical learning generalizes across talkers and, to some extent, new items but these studies involved listeners' active engagement in processing statistics-bearing stimuli. In this study, we manipulated the relationship between voice onset time (VOT) and fundamental frequency (F0) to establish distributional regularities either aligned with American English or reversed to create a subtle foreign accent. We then tested whether statistical learning across passive exposure to these distributions generalized to new items never experienced in the accent. Experiment 1 showed statistical learning across passive exposure but no generalization of learning when exposure and test items shared the same initial consonant but differed in vowel (*bear/pear → beer/pier*) or when they differed in initial consonant but shared distributional regularities across VOT and F0 dimensions (*deer/tear → beer/pier*). Experiment 2 showed generalization to stimuli that shared the statistics-bearing phoneme (*bear/pear → beer/pier*), but only when the response set included tokens from both exposure and generalization stimuli. Moreover, statistical learning transferred to influence the subtle acoustics of listeners' own speech productions but did not generalize to influence productions of stimuli not heard in the accent. In sum, passive exposure is thus sufficient to support statistical learning and its generalization, but task demands modulate this dynamic. Moreover, production does not simply mirror perception: generalization in perception was not accompanied by transfer to production.

Encountering a talker with an idiosyncratic speaking style or a non-native accent can diminish speech comprehension (e.g., Bradlow & Bent, 2008). But experience often leads to improvements that generalize to other contexts (e.g., Xie & Myers 2017). Sometimes, such encounters even impact subtle characteristics of one's own speech (e.g., Pardo et al., 2017). Although instances of such adaptation and convergence are well documented, many questions regarding their bases remain unanswered.

A literature examining dimension-based statistical learning provides a means with which to fill these gaps (e.g., Idemaru & Holt, 2011; Liu & Holt, 2015; Schertz, Cho, Lotto, & Warner, 2016; Wu & Holt, 2022). This work posits that subtle differences across talkers can be characterized as shifts in the underlying acoustic regularities – the statistical distributions – of speech. The somewhat different speech patterns of American English compared to Scottish English (Escudero, 2001), for example, can be modeled as distribution shifts across multidimensional acoustic space, and the impact of listening across these distributions on perception (as well as production) can be tracked.

Such distributional shifts can be studied experimentally. For example, Idemaru and Holt (2011) selectively sample *beer-pier* utterances across an acoustic space defined by voice onset time (VOT, the timing of articulators' release versus voicing onset) and fundamental frequency (F0, related to pitch). A Canonical sampling mirrors the F0xVOT distributions typical of American English: utterances with short VOT tend to have low F0 and be heard as /b/ whereas those with long VOT tend to have higher F0 and be heard as /p/. American English adults' perception mirrors these regularities, with VOT serving as a strong cue to /b/-/p/ category identity and F0 contributing to a lesser extent (Wu & Holt, 2022). Reversing this correlation creates a subtle accent. In a passive exposure version of the paradigm (Hodson et al., 2023; Murphy et al. 2024), listeners hear a sequence of *beer* and *pier* utterances conveying one of these distributional regularities followed by one of two F0-differentiated test stimuli with ambiguous VOT. With only F0 available to convey category identity, test stimulus categorization indexes listeners' reliance on F0 in speech categorization. In a pattern now well-replicated across many studies, F0 robustly signals *beer* versus *pier* when distributions mirror American English norms but F0

reliance is markedly reduced in the context of the accent (Holt, 2025). This points to implicit learning of statistical speech regularities that has an immediate influence on the mapping of acoustics to speech, thus informing how listeners adapt to idiosyncratic or accented speech.

Generalization has been a valuable tool in examining the grain of representation across which this learning operates. For example, if learning operates across talker-specific representations, there should be no generalization to talkers not experienced in the accent. However, learning does generalize to new talkers (Liu & Holt, 2015). Likewise, generalization is evident across lexical items. For example, Idemaru and Holt (2020) report generalization across word contexts with differing vowels (*beer-pier→bear-pear*, and *vice versa*) and differing vowel-consonant frames (*beer-pier→bill-pill)*, although generalization effects were weaker than effects for the token experienced in the accent (see also Liu & Holt, 2015; Lehet & Holt, 2020; Zhang, Wu & Holt, 2021). In contrast, generalization is not apparent across the acoustic dimensions that convey speech regularities, like F0 and VOT. Idemaru and Holt (2014) find that *beer-pier* learning does not generalize to influence *deer-tear* although each samples a similar F0xVOT acoustic space.

Collectively, these studies point to phoneme-sensitive learning. However, in contrast to the passive exposure paradigm described above (Hodson et al., 2023; Murphy et al. 2024), generalization studies have relied exclusively on active tasks with overt, trial-by-trial categorization of both statistics-bearing "exposure" speech stimuli and the "test" stimuli that measure statistical learning and subsequent generalization. Correspondingly, in these prior studies, the response set includes responses that match the statistics-bearing speech (e.g., *bear-pear*) as well as responses to test generalization (*beer-pier*). This might be important. Wu and Holt (2022) observe that individual differences in the strength of category activation – as indexed by categorization accuracy for statistics-bearing exposure stimuli – predict the magnitude of down-weighting of F0 upon introduction of the accent. If active categorization were to more robustly drive category activation than mere exposure, there may be task-driven learning and/or generalization outcomes. Hodson et al. (2023) examined this

possibility, finding common patterns of F0 down-weighting for active trial-by-trial categorization of exposure stimuli and passive exposure to them. Yet there remains an open question: is active categorization across statistics-bearing stimuli necessary for generalization? This paper tackles this question.

Experiment 1 examines generalization of statistical learning across passive exposure with a single response set (*beer-pier)* across all conditions. Listeners hear a sequence of utterances conveying canonical or reverse distributions, then categorize sequence-final, F0-differentiated *beer-pier* test stimuli across three conditions: No Generalization (*beer-pier→beer-pier*), same Phoneme Generalization (*bear-pear→beer-pier,* for which active categorization paradigms observe generalization), and same Dimension Generalization (*deer-tear→beer-pier,* for which no generalization is observed in active tasks). To anticipate the results, we replicate the null effect in the Dimension Generalization condition. But, unlike past studies, we do not observe generalization in the Phoneme Generalization condition in Experiment 1. In Experiment 2 we examine whether this difference arises from learning differences across passive exposure. We focus our investigation on exposure and test stimuli that share initial and final phonemes and introduce a mixed response set (*beer-pier + bear-pear)* in the critical condition. The mixed response set restores generalization, despite the passive exposure.

As a secondary measure in each experiment, we elicit speech productions to attempt to replicate recently reported transfer of statistical learning from perception to production (Murphy et al., 2024) and to examine generalization in production. To foreshadow, we robustly replicate Murphy and colleagues (2024) for statistics-bearing stimuli: the learning arising with perceptual experience with an accent transfers to impact listeners' speech productions. Intriguingly, this transfer is limited to stimuli heard in the accent. Generalization in perception is not reflected in production.

Experiment 1

Methods

Experiment 1 examined statistical learning across passive exposure to speech regularities and its generalization. Participants listened to a sequence of speech tokens possessing a (Canonical, Reverse) short-term distributional regularity and reported whether a final test stimulus was *beer* or *pier.* They then heard the same test stimulus again and repeated it aloud (Figure 1). Test stimuli were always *beer-pier*, differentiated only by F0. Across conditions experienced by all listeners, the stimuli that conveyed distributional regularities across passive exposure varied: *beer-pier* (requiring No Generalization), *bear-pear* (Phoneme Generalization), *deer-tear* (Dimension Generalization).

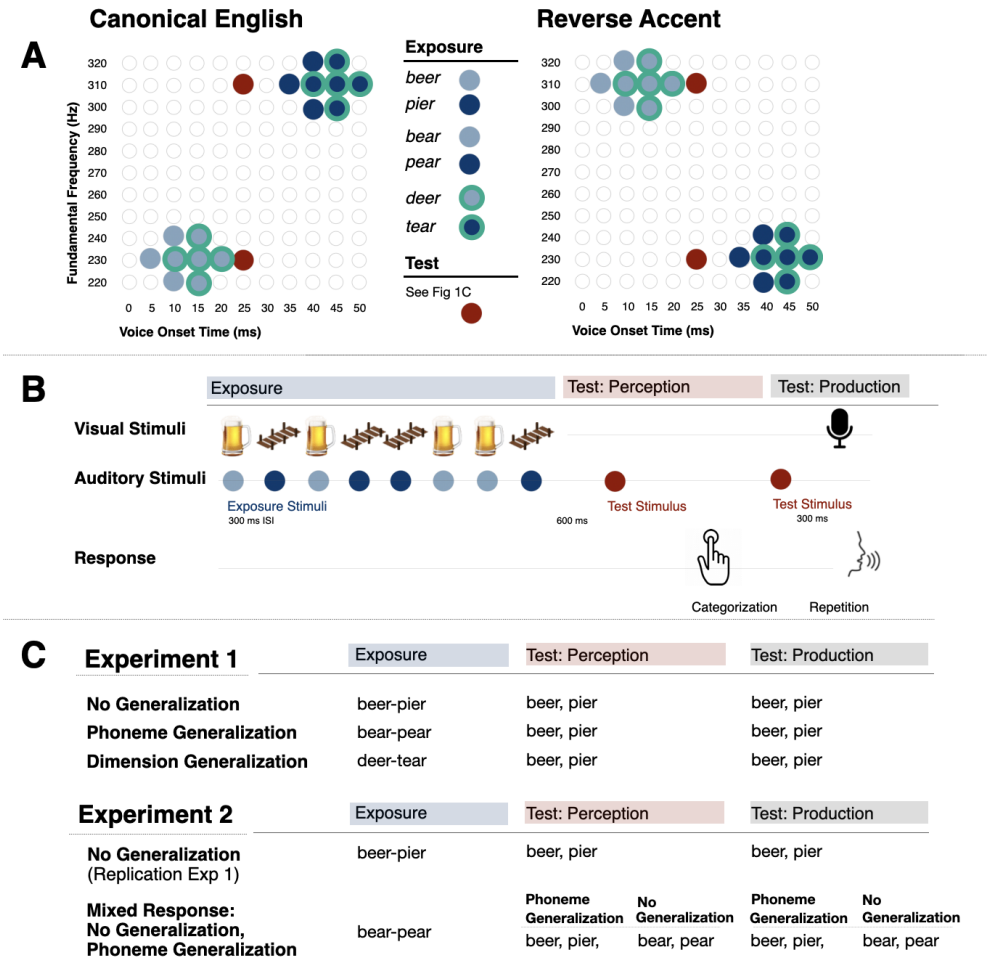

Figure 1. Experiment Protocol. A. Stimuli. An acoustic space defined by voice onset time (VOT) and fundamental frequency (F0) conveyed *beer-pier* and *bear-pear* (solid blue, no line) and *deer-tear* (solid blue, aqua line) tokens sampled in a manner Canonical of American English or Reversed to convey an accent. **B. Trial Structure.** A representative trial from the Experiment 1 Control condition illustrates the trial structure across each experiment, and all groups. **C. Experiment Conditions.** The speech tokens that convey the short-term speech regularity (Exposure, blue) and the test stimuli that elicit perception (Perception, red) and production (Production, gray) are depicted for each condition of each experiment.

Participants

In keeping with past studies, we assumed a small effect size of d=0.3 for generalization in speech perception (Liu and Holt 2015; Idemaru and Holt, 2020). A power analysis performed using the program PANGEA (Westfall, 2015) indicated that a sample size of 90 participants would provide power > 0.8 to detect a three-way interaction between Test Stimulus F0, Canonical/Reverse statistical regularity and the three-level generalization factor, at $\alpha = 0.05$. As a provision against data loss in online studies, we collected online data from 110 adult (55 females) native-English participants located in the United States. Eighteen participants' data did not enter into analyses due to silent or highly noisy production

recordings that precluded acoustic analysis of speech productions (N=17) or perceptual responses indicating task noncompliance (N=1). Data from 92 participants (49 females, mean age 28.1 years, SD = 4.8 years) entered the analysis.

Stimuli

Figure 1A illustrates the speech stimuli. Fundamental frequency (F0) and voice onset time (VOT) varied, with other acoustic dimensions held constant, to create perceptual spaces corresponding to *beer-pier, bear-pear*, and *deer-tear.* Each of the six target words was spoken by an adult female native American English speaker, with specific tokens chosen to have similar duration (400 ms for *beer-pier* and *deer/tear*, 500 ms for *bear-pear*) and F0 contour. Beginning with these natural speech exemplars, we edited in the time domain to create 5-ms VOT steps (McMurray & Aslin, 2005). Next, we manipulated the F0 onset of each of these stimuli using a custom Praat script (Praat 6.1, Boersma & Weenink, 2023) such that onset F0 varied from 220 to 320 Hz in 10 Hz steps, with F0 contour interpolated smoothly across voicing to word offset. Amplitude normalization assured each stimulus possessed the same root mean-squared amplitude.

Exposure stimuli (blue, Figure 1A) sub-sampled these acoustic spaces to create distinct short-term speech regularities. The Canonical English sampling (Figure 1A, left) followed acoustic speech regularities typical of American English: stimuli with shorter VOT (<25 ms) tend to have lower F0 and be labeled as /b/ or /d/ (light blue) whereas those with longer VOT (>25 ms) tend to have higher F0 and be labeled as /p/ or /t/ (dark blue). A statistically defined ‘accent' reversed this distributional relationship from American English norms (Figure 1A, right). Here, for the Reverse condition, shorter VOTs signal /b/ or /d/ but F0 is *higher* frequency. Longer VOTs signal /p/ or /t/ but F0 is *lower* frequency. *Beer-pier* and *bear-pear* tokens (blue, no line) shared identical F0xVOT values whereas *deer-tear* tokens (blue, aqua line) sampled distributions shifted +5 ms in VOT to account for natural English VOT patterns (Cho & Ladefoged, 1999).

Additionally, two test stimuli possessed a perceptually ambiguous, 25-ms VOT and varied only in F0 (230 or 310 Hz; Figure 1A, red symbols). Test stimulus categorization measured listeners' reliance on F0 in category decisions related to learning (when test stimuli match exposure stimuli, e.g., *beer-pier→beer-pier*) and generalization (*bear-pear→beer-pier* and *deer-tear→beer-pier*). These same test stimuli elicited speech productions in the auditory repetition task. Exposure and Test stimuli were chosen on the basis of responses provided by nine raters and had been previously shown to drive statistical learning in perception (Murphy, 2024).

Procedure

Online participants recruited via Prolific.co were automatically directed to an experiment hosted on Gorilla (www.gorilla.sc, Anwyl-Irvine et al., 2021). Using the Chrome browser on a computer (no mobile devices), participants provided consent, completed a demographics survey, and underwent both a brief check of headphone compliance test (Milne et al., 2020) and a check that the computer microphone was recording utterances.

Figure 1B shows the trial structure. Participants listened passively to a sequence of 8 perceptually unambiguous exposure stimuli that conveyed either a Canonical or a Reverse short-term regularity. Each sequence included 4 tokens from each of the two distributions (Figure 1A, dark and light blue symbols), randomly selected and concatenated with 300-ms silent intervals separating utterances. Clipart images corresponding to the word expected from the perceptually unambiguous VOT appeared at the onset of each sound. Next, after 600 ms, participants heard one of the two test stimuli (High or Low F0; Figure 1A, red symbols) and categorized it as *beer* or *pier* via a keyboard response with onscreen text to guide the mapping. Then, 300 ms later, the same test stimulus played again, and an image of a microphone prompted participants to repeat the word aloud. Participants had 2500 ms to repeat the test stimulus and utterances were saved digitally for subsequent acoustic analysis of F0.

As summarized in Figure 1C, *beer-pier* test stimuli elicited perceptual categorization responses and speech productions across each of three conditions. The statistics-bearing exposure stimuli of the No Generalization condition matched the *beer-pier* test stimuli, thereby measuring statistical learning without requiring generalization. In contrast, *bear-pear* exposure sequences in the Phoneme Generalization condition necessitated generalization of statistical learning to *beer-pier* test stimuli sharing a common initial phoneme. Finally, in the Dimension Generalization condition, the *deer-tear* regularities differed in initial phoneme from the *beer-pier* test stimuli but overlapped across F0xVOT acoustic dimensions.

For each condition, participants experienced 30 Canonical trials followed by 30 Reverse trials. Among these, 24 of 30 stimuli involved exposure stimuli followed by one of the two *beer-pier* VOT-ambiguous, F0-differentiated test stimuli described above; responses to these stimuli entered analyses. Six additional VOT-unambiguous stimuli served as a data-quality check of online participants, with *a priori* exclusion of participants who gave the same response to these unambiguous stimuli (no participants were excluded on this basis). Unambiguous *beer, bear,* and *deer* stimuli had a 230 Hz F0 and 10ms VOT (*deer*: 15ms VOT) while unambiguous *pier, pear,* and *tear* stimuli has a 310 Hz F0 and 40ms VOT (tear: 45ms VOT). As in prior studies (Wu & Holt, 2022), categorization of these perceptually unambiguous stimuli was consistent with expectations from English (Long VOT, 96% /p/; Short VOT, 93% /b/). A Latin square design assured balanced presentation of conditions across participants.

Statistical Analyses

Perceptual Categorization. We modeled the influence of statistical learning on perceptual categorization of test stimuli using mixed effects models (*lme4*, Bates, Mochler, Bolker, and Walker, 2015) in R (version 4.1.3, R Core Development Team, 2022) with the binary (*beer, pier*) categorization response as the dependent variable. The full statistical model involved fixed effects across Statistical Regularity (Canonical, Reverse), Test Stimulus F0 (Low F0, High F0) and Condition (No Generalization,

beer-pier; Phoneme Generalization, *bear-pear*; Dimension Generalization, *deer-tear*) as well as 2- and 3-way interactions. Random effects included by-subject random intercepts and random slopes for Statistical Regularity and Test Stimulus F0 over subjects. Statistical Regularity and Test Stimulus F0 fixed effects were center coded (-0.5 or 0.5). A simple effects coding scheme was applied to the 3-level Condition effect whereby the No Generalization condition served as the reference level to which the Phoneme Generalization and Dimension Generalization conditions were compared. Three-way interactions among Statistical Regularity, Test Stimulus F0, and Condition were examined with post-hoc tests of the Statistical Regularity by Test Stimulus F0 interaction for each Condition. Satterthwaite approximates using the *LmerTest* package (version 3.1-3, Kuznetsova, Brockhoff, & Christensen, 2016) provided *p* values.

Speech Production. Transfer of statistical learning in listening to repetition productions was modeled across by-participant z-score normalized utterance F0 (as in Murphy et al., 2024). In brief, the F0 (computed across the first 40 ms) was measured for each utterance. F0 values ±3 standard deviations from a participant's mean F0 were removed from analysis. Next, we normalized F0 on a by-individual basis to account for F0 variability arising across talkers (Titze, 1989). Therefore, for production analyses, a z-score of 0 indicates the mean F0 for a participant across all productions. Positive and negative z-scores correspond to continuous standard deviation units above and below the mean, respectively, that we submitted to standard linear effects models. Fixed and random effect structures, and the approach to post-hoc tests, were identical to perceptual statistical learning analyses.

Results

Perceptual Categorization

Figure 2 presents perceptual categorization of F0-differentiated *beer-pier* test stimuli as a function in Canonical and Reverse conditions. Table 1 displays results of a logistic mixed effects model fit to these data.

Across all conditions, there were more *pier* responses for High F0, as is typical in American English (Lisker, 1986), reflected in a main effect of Test Stimulus F0 ($z=17.52$, $p<.001$) and a main effect of Statistical Regularity ($z=17.53$, $p<.001$). Importantly, these factors significantly interacted ($z=15.43$, $p<.001$), indicating that statistical learning across passive listening impacted reliance on F0 in categorization.

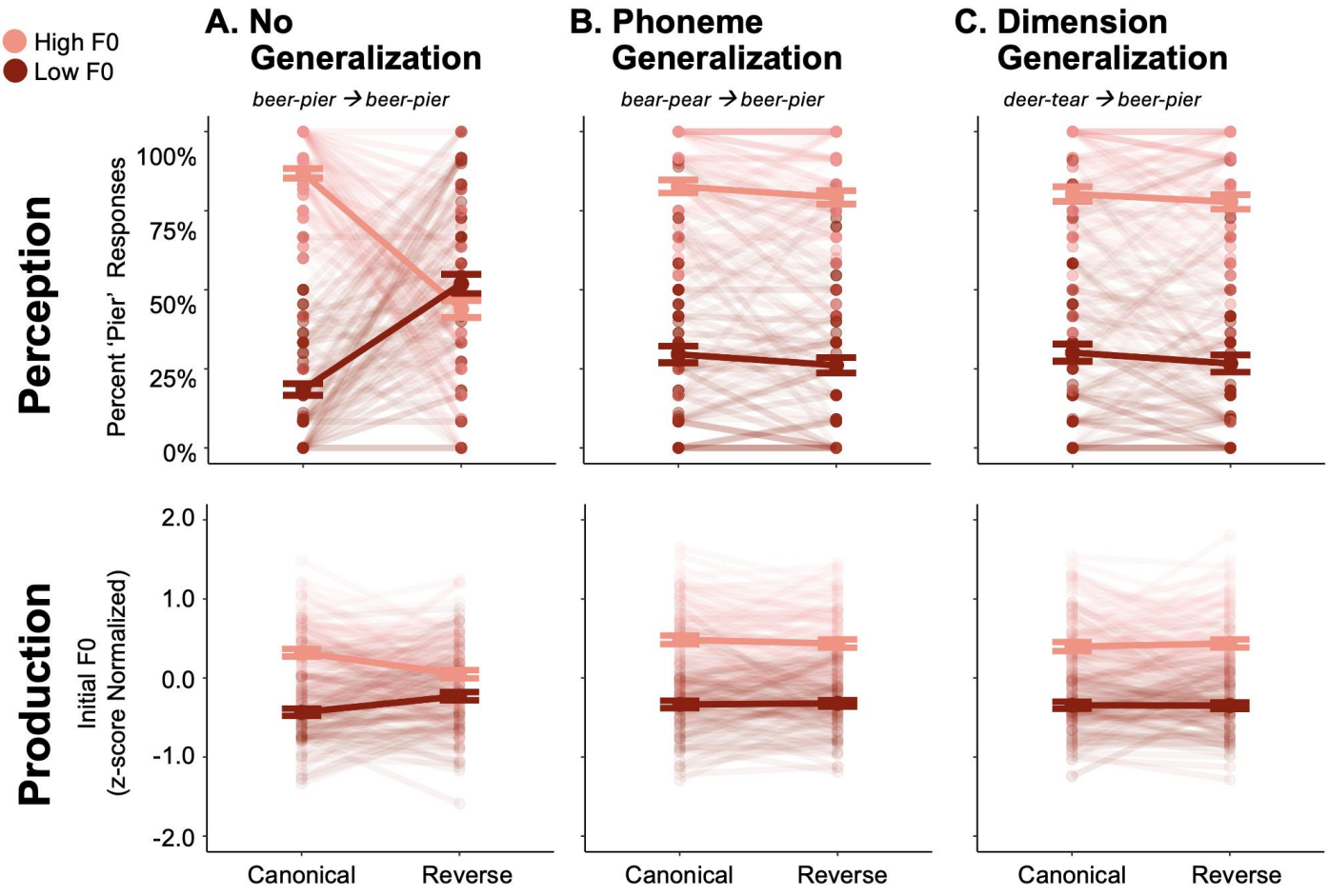

Figure 2. Experiment 1 Perception and Production Results. The top row depicts percent *pier* categorization responses to High and Low F0 *beer-pier* test stimuli in the context of Canonical and Reverse short-term regularities. The bottom row shows z-score normalized fundamental frequency (F0) of *beer-pier* speech productions elicited in repetition of High and Low F0 test stimuli in the context of Canonical and Reverse short-term regularities. **A.** No Generalization (*beer-pier* exposure, *beer-pier* test) **B.** Phoneme Generalization (*bear-pear* exposure, *beer-pier* test). **C.** Dimension Generalization (*deer-tear* exposure, *beer-pier* test). Larger symbols and thick lines represent sample mean and standard error. Smaller symbols and transparent lines indicate individual participants' behavior.

Simple effects coding comparing perceptual responses from the No Generalization condition to responses from the Phoneme Generalization and Dimension Generalization revealed significant main effects (Phoneme: z=4.22, p<.001; Dimension: z=3.17, p=.002) indicative of an overall difference in perceptual response across conditions. Two-way interactions were significant between each condition and Test Stimulus F0 (Test Stimulus F0 by Phoneme: z=8.97, p<.001; Test Stimulus F0 by Dimension: z=7.48, p<.001) but not Statistical Regularity. Importantly, two significant three-way interactions indicated that perceptual down-weighting differs for both the Phoneme Generalization (z=-18.70, p<.001), and Dimension Generalization (z=-19.27, p<.001) conditions relative to the No Generalization condition.

Based on these two significant three-way interactions, we tested statistical learning/generalization in each condition with separate, post-hoc logistic mixed effect models. The two-way interaction between Test Stimulus F0 and Statistical Regularity was significant only in the No Generalization model (z=23.72, p<.001), but not the Phoneme (z=0.57, p=.567) or Dimension (z=0.94, p=.348) Generalization models. Thus, Experiment 1 reveals evidence of statistical learning but not of generalization of the learning.

Table 1. Experiment 1 Perceptual Categorization of Test Stimuli across Conditions

	β	*SE*	*z*	*p*
Intercept	0.22	0.08	2.74	.006
Statistical Regularity	0.24	0.05	4.48	<.001
Test Stimulus F0	2.40	0.14	17.53	<.001
Phoneme Generalization	0.24	0.06	4.22	<.001
Dimension Generalization	0.18	0.06	3.17	.002
Statistical Regularity x Test Stimulus F0	1.45	0.09	15.43	<.001
Statistical Regularity x Phoneme Generalization	-0.10	0.11	-0.85	.396
Statistical Regularity x Dimension Generalization	-0.14	0.11	-1.28	.199
Test stimulus F0 x Phoneme Generalization	1.01	0.11	8.97	<.001
Test stimulus F0 x Dimension Generalization	0.83	0.11	7.48	<.001
Statistical Regularity x Test Stimulus F0 x Phoneme Generalization	-4.20	0.22	-18.72	<.001
Statistical Regularity x Test Stimulus F0 x Dimension Generalization	-4.28	0.22	-19.27	<.001

Note: Reference levels are Statistical Regularity (Reverse), Test Stimulus F0 (Low F0), Condition (No Generalization). Phoneme Generalization and Dimension Generalization result from simple effects coding comparing the respective conditions to the No Generalization condition.

Speech Production

Figure 2 shows z-score normalized F0 measured from participants' *beer-pier* speech productions as a function of the Statistical Regularity. Table 2 provides results of the Linear Mixed Effects Model.

Overall, speech productions elicited by the High (compared to the Low) F0 *beer-pier* test stimuli had higher F0 ($t=15.35$, $p<.001$). A significant two-way interaction between Test Stimulus F0 and Statistical Regularity ($z=5.11$, $p<.001$) indicated transfer of statistical learning to production. Simple effects coding comparing production F0s from the No Generalization Condition to production F0s from Phoneme Generalization and Dimension Generalization revealed significant main effects of each Condition (Phoneme: $z=7.05$, $p<.001$; Dimension: $z=5.10$, $p<.001$). Significant two-way interactions were also evident between each Condition and Test Stimulus F0 (Test Stimulus F0 by Phoneme: $t=6.56$, $p<.001$; Test Stimulus F0 by Dimension: $t=6.19$, $p<.001$). As with the perceptual categorization results, two significant three-way interactions indicated that transfer of statistical learning to production differed in both the Phoneme Generalization ($t=-5.65$, $p<.001$) and Dimension Generalization ($t=-6.74$, $p<.001$) Conditions relative to the No Generalization Condition. Also similar to perception, the post-hoc analyses only revealed a significant interaction between Test Stimulus F0 and Statistical Regularity in the No Generalization condition ($t=9.44$, $p<.001$)[1], but not in the Phoneme Generalization ($t=0.87$, $p=.383$) or Dimension Generalization ($t=-0.82$, $p=.410$) condition.

The results are clear: perceptual statistical learning across passive exposure failed to generalize in perception. While this replicates the finding of no generalization in the Dimension Generalization (*deer-tear→beer-pier*) condition (Idemaru & Holt, 2014), it contrasts with Phoneme Generalization (*bear-pear→beer-pier*) observed in active tasks that involve trial-by-trial overt speech categorization

(Idemaru & Holt, 2020). Naturally, since no generalization was uncovered in perception, transfer of generalization was not seen in production.

Table 2. Experiment 1 Speech Production F0 across Conditions

	β	*SE*	*t*	*p*
Intercept	0.01	0.01	1.03	.302
Test Stimulus F0	0.69	0.04	15.35	<.001
Statistical Regularity	0.02	0.03	0.63	.533
Phoneme Generalization	0.14	0.02	7.05	<.001
Dimension Generalization	0.10	0.02	5.10	<.001
Test Stimulus F0 x Statistical Regularity	0.16	0.03	5.11	<.001
Test Stimulus F0 x Phoneme Generalization	0.26	0.04	6.56	<.001
Test Stimulus F0 x Dimension Generalization	0.24	0.04	6.19	<.001
Statistical Regularity x Phoneme Generalization	-0.02	0.04	-0.47	.641
Statistical Regularity x Dimension Generalization	-0.06	0.04	-1.45	.148
Test Stimulus F0xStatistical Regularity x Phoneme Generalization	-0.44	0.08	-5.65	<.001
Test Stimulus F0 x Statistical Regularity x Dimension Generalization	-0.53	0.08	-6.74	<.001

Note: Reference levels are Statistical Regularity (Reverse), Test Stimulus F0 (Low F0), Condition (No Generalization). Phoneme Generalization and Dimension Generalization result from simple effects coding comparing the respective conditions to the No Generalization condition.

Experiment 2

Methods

Experiment 1 replicated the null effect of dimension generalization (Idemaru & Holt, 2014) but failed to find evidence of phoneme generalization, contrary to prior reports (Idemaru & Holt, 2020). One interpretation of these results is that statistical learning across passive listening is not sufficient to support generalization. But before this conclusion is drawn, we must rule out the influence of another factor. Recall that in Idemaru and Holt's (2020) task, participants responded to all tokens, meaning that both the statistics-bearing stimuli and the generalization stimuli were part of the response set. If overlap between exposure and test stimuli is critical for extracting statistics or applying statistics to new stimuli, then a mixed response set should restore phoneme generalization, even with passive exposure. Experiment 2 tested this possibility. First, we aimed to replicate the main findings of statistical learning and its transfer to production, observed in Experiment 1, in a different pair, *bear-pear.* We used this

pair as exposure stimuli to test phoneme generalization to a different pair, *beer-pier,* presented in a mixed response set comprised of both *bear-pear* and *beer-pier* tokens with equal frequency.

Participants

Based on the power analysis of Experiment 1, we tested 95 participants (48 female) with 87 participants (45 female, M_{age} = 31.3, SD = 6.0 years) entering analyses after application of the Experiment 1 exclusion criteria.

Stimuli

Experiment 2 relied on the *beer-pier* and *bear-pear* stimuli from Experiment 1 (Figure 1A).

Procedure

Experiment 2 consisted of 6 blocks (30 trials each) of trials alternating with Canonical and Reverse regularities. The first two blocks reproduced the No Generalization (*beer-pier→beer-pier*) condition of Experiment 1 (Replication: No Generalization). The remaining four blocks conveyed statistics across *bear-pear* utterances and involved both *bear-pear* (Mixed Response: No Generalization) and *beer-pier* (Mixed Response: Phoneme Generalization) test trials, randomly intermixed such that there was uncertainty about the target of categorization on each trial and the mixed response set involved *beer, pier, bear,* and *pear*. As in Experiment 1, in each block six VOT-unambiguous trials assured online participants' data quality; no participants were excluded on the basis of responses to these trials. Performance was high and consistent with English regularities (Long VOT, 93% /p/; Short VOT, 88% /b/). Responses from these trials did not enter analyses, resulting in 24 Canonical and 24 Reverse trials for each condition.

Statistical Analyses

Perceptual Categorization. The statistical approach was similar to Experiment 1. Our first goal was to replicate statistical learning and its transfer to production in the No Generalization condition, observed in Exp 1. This model included the subset of data from the *beer-pier→beer-pier blocks.* The model included Test Stimulus F0 (High F0, Low F0), Statistical Regularity (Canonical, Reverse) and their interaction, as well as a maximal random effects structure consisting of by-subject random intercept, random slopes for Test Stimulus F0, Statistical Regularity, and the interaction between Test Stimulus F0 and Statistical Regularity over subjects. As in Experiment 1, Statistical Regularity and Test Stimulus F0 fixed effects were center coded (-0.5 or 0.5).

Next, we examined generalization using blocks with Mixed Response conditions. The model's dependent variable was coded as voiced (beer, bear) or voiceless (pier, pear). Three fixed effects, Test Stimulus F0 (High F0, Low F0), Statistical Regularity (Canonical, Reverse) and Condition (Mixed Response: No Generalization; Mixed Response: Phoneme Generalization), were included alongside their 2-way and 3-way interactions. The random effects structure was similar to the structure used in the Replication task analysis with the addition of a random slope for Condition. All fixed effects were centered coded (-0.5, or 0.5).

Speech Production. Acoustic speech analysis followed the Experiment 1 approach with by-participant z-score normalized production F0s as a continuous dependent variable analyzed with linear mixed effects models. As with the perceptual categorization analysis, separate models assessed production changes in the Replication and the Mixed Response tasks. Fixed effects and their interactions were identical to those included in the corresponding perceptual categorization models. Both models included by-participant random intercept and random slopes for Test Stimulus F0 and Statistical Regularity. The Mixed Response model also included a random slope for Condition. Neither model tolerated the addition of random slopes for the interaction terms. All fixed effects were center coded (-0.5 or 0.5).

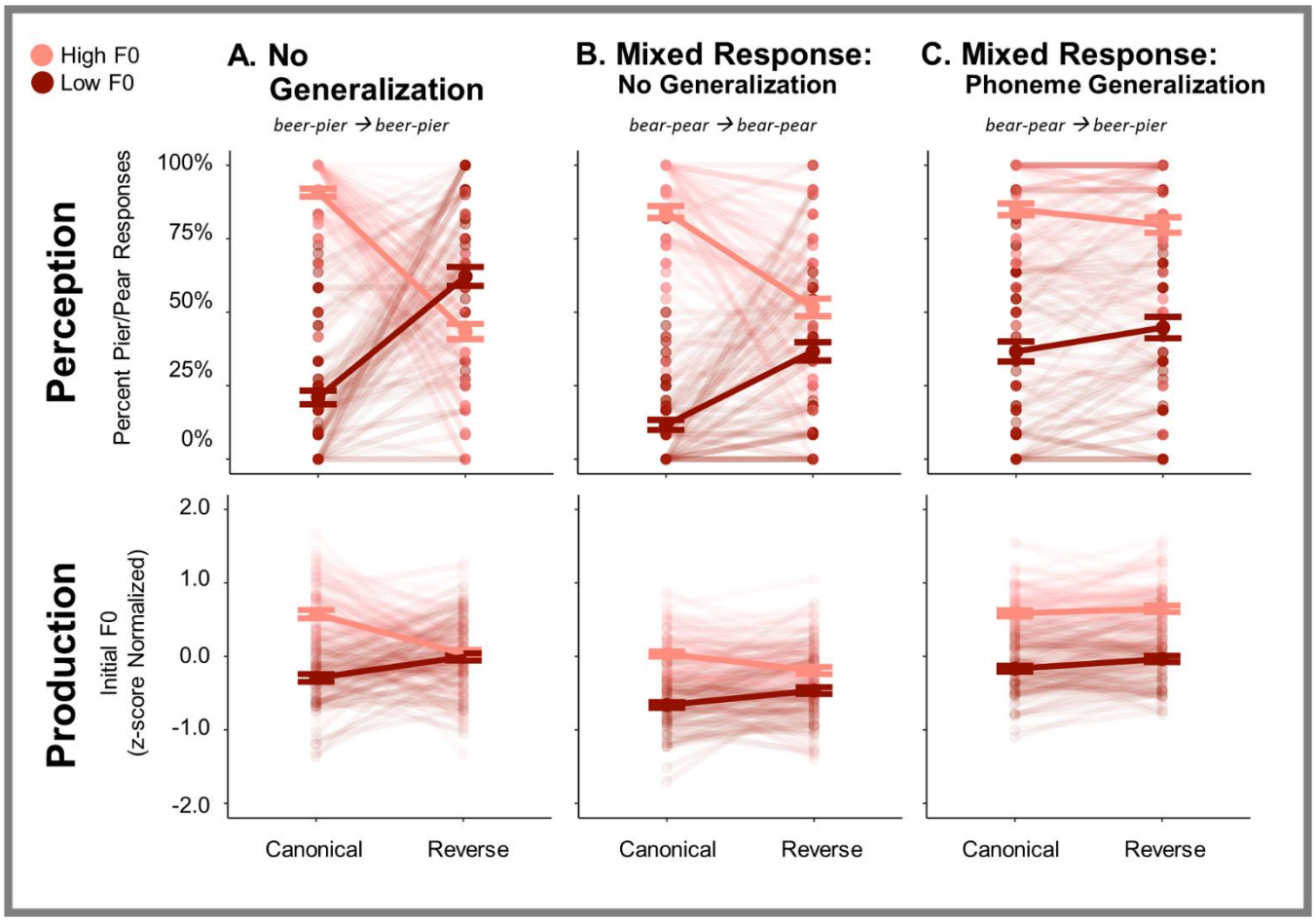

Figure 3. Experiment 2 Perception and Production Results. The top row depicts percent *pier/pear* categorization responses to High and Low F0 *beer-pier* (**A**, **C**) or *bear-pear* (**B**) test stimuli in the context of Canonical and Reverse short-term regularities. The bottom row shows z-score normalized fundamental frequency (F0) speech productions elicited in repetition of these same test stimuli. **A.** Replication: No Generalization (*beer-pier* exposure, *beer-pier* test) is a replication of Experiment 1. **B.** Mixed Response Condition trials with No Generalization (*bear-pear* exposure, *bear-pear* test). **C.** Mixed Response Condition trials requiring Phoneme Generalization (*bear-pear* exposure, *beer-pier* test).

Results

Perceptual Categorization

As in Experiment 1, we analyzed perceptual responses for evidence of statistical learning and its generalization to novel tokens (Figure 3, top row). Analysis of perceptual responses from the Replication task revealed a significant main effect of Test Stimulus F0 ($z=8.18$, $p<.001$), a significant main effect of Statistical Regularity ($z=2.00$, $p=.045$) and, importantly, an interaction between the two ($z=13.87$, $p<.001$), showing statistical learning in perception.

Table 3 reports the results from the analysis of perceptual response from the Mixed Response blocks. A significant main effect of Test Stimulus F0 indicated that, overall, participants tended to perceive High F0 test stimuli as *pier or pear* and Low F0 as *beer or bear* (z=18.58, p<.001). The main effect of Condition was also significant, indicating a difference in /b/ versus /p/ response rates in the Mixed Response: No Generalization and the Mixed Response: Phoneme Generalization conditions (z= -5.70, p<.001). This difference appears to be driven by a bias towards *pier* responses in the Mixed Response: Phoneme Generalization condition, a finding also reported by Idemaru & Holt (2020). A significant two-way interaction between Statistical Regularity and Test Stimulus F0 indicated statistical learning in perception in the Mixed Response blocks (z= 12.50, p<.001).

There was also a significant three-way interaction between Statistical Regularity, Test Stimulus F0, and Condition (z=10.20, p<.001). To unpack this interaction, we fit separate post-hoc models to each of the two Conditions, separately. In the Mixed Response blocks, there is evidence of statistical learning in the form of a significant two-way interaction between Statistical Regularity and Test Stimulus F0 in both the No Generalization model (z=11.47, p<.001) as well as the Phoneme-Generalization model (z=3.49, p<.001).

Table 3. Perceptual Categorization of Voiced/Voiceless Test Stimuli in Mixed Response Task

	β	*SE*	*z*	*p*
(Intercept)	0.26	0.12	2.18	.029
Statistical Regularity	0.02	0.08	0.21	.833
Test Stimulus F0	2.78	0.15	18.58	<.001
Condition	-1.16	0.20	-5.70	<.001
Statistical Regularity x Test Stimulus F0	2.23	0.18	12.50	<.001
Statistical Regularity x Condition	0.18	0.13	1.38	.168
Test Stimulus F0 x Condition	-0.18	0.14	-1.23	.218
Statistical Regularity x Test Stimulus F0 x Condition	2.72	0.27	10.20	<.001

Note: Reference levels are Statistical Regularity (Reverse), Target stimulus F0 (Low F0), Condition (Phoneme-Generalization)

Speech Production

We next examined transfer of statistical learning to production using z-score normalized F0 measured from *beer-pier* and *bear-pear* productions (Figure 3, bottom row). First examining the Replication condition, the model reveals the expected main effect of Test Stimulus F0 ($t=9.55$, $p<.001$), as well as a significant two-way interaction between Test Stimulus F0 and Statistical Regularity indicating the transfer of statistical learning to production ($t=14.55$, $p<.001$), thereby replicating the transfer observed in Experiment 1[2].

Table 4 reports the transfer of speech production results from the Mixed Response blocks. Mirroring the perceptual results, the analysis revealed a main effect of Test Stimulus F0 ($t = 13.64$, $p <.001$), as well as a main effect of Condition on production F0s ($t=-14.81$, $p<.001$). The latter finding is in line with previous research on intrinsic F0, a tendency for high vowels like the /i/ in *beer* to have higher F0s than low vowels like the /e/ in *bear* (Chen et al. 2021; Whalen & Levitt, 1995). Transfer of statistical learning was evident in the significant two-way interaction between Statistical Regularity and Test Stimulus F0 ($t=6.64$, $p<.001$)[3]. There were significant interactions between Statistical Regularity and Condition ($t=3.01$, $p=.003$) as well as Test Stimulus F0 and Condition ($t=-6.41$, $p<.001$).

Critical for our determining whether generalization transfers to influence speech production, we found a significant three-way interaction between Statistical Regularity, Test Stimulus F0, and Condition ($t=4.63$, $p<.001$). Post hoc analyses revealed that the two-way interaction between Test Stimulus F0 and Statistical Regularity was significant in the Mixed Response: No Generalization model ($t=8.18$, $p<.001$) but the perceptual generalization observed for the Mixed Response: Phoneme-Generalization condition did not transfer to production ($t=1.44$, $p=.151$).

To summarize, Experiment 2 replicates statistical learning across passive exposure to *beer-pier* and its transfer to speech production. It extends this finding to *bear-pear*, when no generalization is required. Importantly, inclusion of a mixed response set rescued phoneme-level generalization of perceptual statistical learning, although with a smaller magnitude of influence on the generalization pair than the pair experienced across the regularity. This generalization of learning did not transfer to influence speech production.

Table 4. Mixed Response Task Productions by Test Stimulus F0 and Condition

	β	*SE*	*t*	*p*
(Intercept)	-0.03	0.01	-2.34	0.021
Statistical Regularity	-0.04	0.03	-1.36	0.177
Test Stimulus F0	0.60	0.04	13.64	<.001
Condition	-0.58	0.04	-14.81	<.001
Statistical Regularity x Test Stimulus F0	0.24	0.04	6.64	<.001
Statistical Regularity x Condition	0.11	0.04	3.01	0.003
Test Stimulus F0 x Condition	-0.24	0.04	-6.41	<.001
Statistical Regularity x Test Stimulus F0 x Condition	0.34	0.07	4.62	<.001

Note: Reference levels are Statistical Regularity (Reverse), Target stimulus F0 (Low F0), Condition (Phoneme-Generalization)

General Discussion

Does generalization of statistical learning emerge only with learning in an active task? Potentially consistent with this possibility, Wu and Holt (2022) argued that when speech conveys sufficient perceptual information to activate a phonetic category (e.g., via unambiguous VOT) it may generate predictions of the typical mapping of other secondarily diagnostic acoustic dimensions, like F0, to the category representation. In the Reverse condition, these expectations are not met and the mismatch may power error-driven learning that down-weights F0 to minimize future mismatches. Inasmuch as active engagement in a categorization decision might boost category activation, it thus may promote learning and its successful generalization. Yet, Hodson et al. (2023) report statistically equivalent learning outcomes across passive exposure to statistics-bearing speech stimuli and active engagement

in a categorization decision across these same stimuli. This latter result suggests that learning across passive exposure may be just as potent as learning across stimuli that demand active categorization. Experiment 2 confirms that statistical learning across passive listening is sufficient to support generalization to stimuli never heard in the accent. Notably, this pattern was not evident in Experiment 1. The difference was that in Experiment 2, a response set included both statistics-bearing and new stimuli with the same initial phoneme. This restored generalization of the learning that accrued across passive listening to the accent.

But why should response set matter? Although speculative, the most reasonable explanation for the influence of response set on generalization may relate to attention and goal-setting, in line with recent findings that show the importance of explicit attentional goals in implicit statistical learning (Zhang & Carlisle, 2023). If participants detect no relationship between exposure and test stimuli, they may tune out exposure stimuli. Under this view, attention is important for learning not because it forces the learner to actively process each statistics-bearing stimulus, but rather because it sets a higher-level behavioral goal in the cognitive-perceptual system. Our results demonstrate the importance of task demands and goals in the context of statistical learning, even when it emerges implicitly across passive exposure. This argues for further research to examine how implicit and explicit task demands influence the nature of information learned across passive exposure.

The present study also lays groundwork for understanding the structure of representation shared between speech perception and production. We replicated the transfer of statistical learning from perception to production reported in Murphy et al. (2024) twice (Experiment 1 and 2, *beer-pier→beer-pier*). Additionally, the present work extends evidence of transfer to a novel word pair (Experiment 2, *bear-pear→bear-pear*). These results demonstrate that there are rapid and implicit changes to the production system as a result of statistical learning across the patterns of other talkers' speech. They are interesting, particularly, in light of the finding that most instances of auditory repetition are carried

out through the “lexical”, as opposed to the “nonlexical” route (Nozari et al., 2010; Nozari & Dell, 2013). This means that upon hearing a word, the individual retrieves the corresponding stored lexical representation and activates the production chain, rather than simply mapping input to output phonology without fully engaging the production system. In our case, the perceptual judgment performed before production makes it even more likely for participants to use the lexical route. Nevertheless, we observe changes to production. This implies that the results do not reflect a simple imitation of the input. As such, the present data build from Murphy et al. (2024) to provide new insights into phonetic convergence (Pardo, 2022) and to extend how other talkers’ speech affects one’s own productions (e.g. Bourguignon et al., 2014; 2016; Lametti et al., 2014).

Yet, even when *bears* affected *beers* in perception, they did not influence production. In Experiment 2, exposure to *bear-pear* distributional regularities led to statistical learning that generalized to *beer-pier* (with a mixed response set). But this learning did not exert an influence on production. The magnitude of generalization (*bear-pear*→*beer-pier*) was smaller than the magnitude of statistical learning across matched trials (*bear-pear*→*bear-pear*) so it is possible that generalization was not robust enough to drive transfer to production. Alternatively, representations subject to learning in perception may differ from those in production, as has been indicated by previous findings that show changes in production can occur independently of changes in perception (Sheldon & Strange, 1982; Kato & Baese-Berk, 2020; Baese-Berk et al., 2024). Future studies of transfer in dimension-based statistical learning are well-poised to address this intriguing possibility because the approach makes it possible to quantify listeners’ and speakers’ detailed reliance on subtle acoustic dimensions, and to manipulate exposure to distributions across them in both passive and active tasks. At this stage, observance of generalization of statistical learning in the absence of transfer to production is important in establishing that production is not simply a mirror of perceptual experience, according with other studies of statistical learning across speech production and perception (e.g., Kittredge & Dell, 2016;

Schwartz et al. 2012). Learning-related adjustments to the representations within the production system appear to be necessary.

In conclusion, passive exposure is sufficient to produce generalization of statistical learning in perception, but subtle task demands affect generalization. Inasmuch as the utility of implicit statistical learning over passive exposure is its ability to impact behavior, this highlights how important it will be to direct research toward better understanding how statistical learning statistical learning supports, and is influenced by, task goals and demands.

Declarations

Funding

This work was supported by funding from the National Science Foundation BCS-1941357 to LH and BCS-2346989 to LH and BN. TM was supported by the Predoctoral Training Program in Behavioral Brain Research (T32GM081760, awarded institutionally to LH and Julie Fiez).

Conflicts of interest/Competing interests

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

Ethics Approval

Approval was obtained by Institutional Review Board at Carnegie Mellon University.

Consent to Participate

Informed consent was obtained from all individual participants in this study.

Consent for publication

Not applicable

Availability of data and materials

The data and tables of the results are available on OSF (https://osf.io/5uqx8/)

Code Availability

R scripts used for statistical analyses are available on OSF (https://osf.io/5uqx8/)

References

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, *53*(4), 1407-1425.

Babel, M. (2010). Dialect convergence and divergence in New Zealand English. *Language in Society 39*. 437–456.

Baese-Berk, M. M., Kapnoula, E. C., & Samuel, A. G. (2024). The relationship of speech perception and speech production: It's complicated. *Psychonomic Bulletin & Review*, 1-17.

Boersma, P. & Weenink, D. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1, retrieved from http://www.praat.org/

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823

Bourguignon, N. J., Baum, S. R., & Shiller, D. M. (2014). Lexical-perceptual integration influences sensorimotor adaptation in speech. *Frontiers in Human Neuroscience*, *8*, 208.

Bourguignon, N. J., Baum, S. R., & Shiller, D. M. (2016). Please say what this word is—Vowel-extrinsic normalization in the sensorimotor control of speech. *Journal of Experimental Psychology: Human Perception and Performance, 42*(7), 1039-1047.

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

Chen, W. R., Whalen, D. H., & Tiede, M. K. (2021). A dual mechanism for intrinsic f0. *Journal of Phonetics*, *87*, 101063.

Cho, T. & Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics*, *27(2)*, 207-229.

Escudero, Paola. "The role of the input in the development of L1 and L2 sound contrasts: Language-specific cue weighting for vowels." *Proceedings of the 25th annual Boston University conference on language development*. Vol. 1. Somerville, MA: Cascadilla Press, 2001.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, *1*.

Hodson, A. J., Shinn-Cunningham, B., & Holt, L. L. (2023). Statistical learning across passive listening adjusts perceptual weights of speech input dimensions. *Cognition, 238*, 105473.

Holt, L.L. (2025). Speech perception is speech learning. *Current Directions in Psychological Science.*

Hyman, L.M. (1975) *Phonology*. Holt, Reinhart, and Winston; New York

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.

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.

Idemaru, K., & Holt, L. L. (2020). Generalization of dimension-based statistical learning. *Attention, Perception, & Psychophysics*, *82*, 1744-1762.

Kato, M., & Baese-Berk, M. M. (2020). The effect of input prompts on the relationship between perception and production of non-native sounds. *Journal of Phonetics*, *79*, 100964.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*(3), 419-454.

Kittredge, A. K., & Dell, G. S. (2016). Learning to speak by listening: Transfer of phonotactics from perception to production. *Journal of Memory and Language*, *89*, 8-22.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1-26.

Lametti, D. R., Krol, S. A., Shiller, D. M., & Ostry, D. J. (2014). Brief periods of auditory perceptual training can determine the sensory targets of speech motor learning. *Psychological Science, 25*(7), 1325-1336.

Lehet, M., & Holt, L. L. (2017). Dimension-based statistical learning affects both speech perception and production. *Cognitive Science*, *41*, 885-912.

Lehet, M. & Holt, L. L. (2020). Nevertheless, it persists: Perceptual recalibration and normalization of speech impact different levels of representation. *Cognition, 202*,104328.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and Speech*, *29*(1), 3-11.

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1783-1789.

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within- category variation in speech perception. *Cognition*, 95(2), B15-B26.

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.

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

Murphy, Timothy (2024). Transfer of Statistical Learning from Speech Perception to Speech Production. Carnegie Mellon University. Thesis. https://doi.org/10.1184/R1/25315957.v1

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, *21*(4), 422-432.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382-2393.

Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception, & Psychophysics*, *72*(8), 2254-2264.

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*(1), 190-197.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*, 637-659.

Pinet, M., & Iverson, P. (2010). Talker-listener accent interactions in speech-in-noise recognition: Effects of prosodic manipulation as a function of language experience. *The Journal of the Acoustical Society of America*, *128*(3), 1357-1365.

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*(4), 421-436.

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2016). Individual differences in perceptual adaptability of foreign sound categories. *Attention, Perception, & Psychophysics*, *78*(1), 355-367.

Schwartz, J. L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. Journal of Neurolinguistics, 25(5), 336-354.Weil, S. A. (2001). *Foreign accented speech: Adaptation and generalization* (Master's thesis, Ohio State University).

Sheldon, A., & Strange, W. (1982). The acquisition of/r/and/l/by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied psycholinguistics*, *3*(3), 243-261.

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics*, *23*(3), 349-366.

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

Wu, Y. C., & Holt, L. L. (2022). Phonetic category activation predicts the direction and magnitude of perceptual adaptation to accented speech. *Journal of Experimental Psychology: Human Perception and Performance, 48(9), 913-925*.

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America*, *143*(4), 2013-2031.

Zhang, Z., & Carlisle, N. B. (2023). Explicit attentional goals unlock implicit spatial statistical learning. *Journal of Experimental Psychology: General*, *152*(8), 2125-2137.

Zhang, X., & Holt, L. L. (2018). Simultaneous tracking of coevolving distributional regularities in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(11), 1760-1779.

Zhang, X., Wu, Y. C., & Holt, L. L. (2021). The learning signal in perceptual tuning of speech: Bottom up versus top-down information. *Cognitive Science*, *45*(3), e12947.