 doi:10.20944/preprints202105.0777.v1

A learning interaction between statistical learning experiments

Peter T. Richtsmeier[a*] and Lisa Goffman[b]

[a] *Communication Sciences and Disorders, Oklahoma State University, Stillwater, OK, USA;*

[b] *Callier Center for Communication Disorders, Behavioral and Brain Sciences, University of Texas at Dallas, Dallas, TX, USA*

Correspondence to: Dr. Peter T. Richtsmeier

042 Social Sciences and Humanities Hall

Stillwater, OK 74078

prichtsmeier@yahoo.com

 doi:10.20944/preprints202105.0777.v1

A learning interaction between statistical learning experiments

Abstract: When participants in a statistical learning paradigm are asked to learn from two incompatible or competing inputs, they often fail to learn from one or both inputs. This study presents the results of two experiments that were both completed by one group of typically developing four-year-old children. One experiment targeted word-medial consonant patterns (phonotactics), whereas the other targeted strong-weak and weak-strong stress patterns (prosody). The order of the experiments was critical for learning outcomes in the phonotactics experiment: When children learned phonotactics first, their production accuracy increased following exposure to a high frequency input. When children learned phonotactics second, however, their production accuracy dropped when they were exposed to the high frequency input. Results from the prosody experiment were inconclusive, with limited evidence of any learning effect. Overall, the results suggest that children may conflate learning experiences, and patterns learned from an initial experimental input compete with patterns in a subsequent experiment. When considering natural language acquisition, the results suggest that an isolated episode of learning may lead to generalizations that are incompatible with later input, and possibly, with larger patterns in the language.

Abstract Word Count: 180

Keywords: statistical learning, experiment interaction, phonology, child speech, language acquisition

 doi:10.20944/preprints202105.0777.v1

Introduction

Statistical learning has for several decades been a highly productive framework for studying the learning of a variety of linguistic structures, as well as structured visual input (for an overview see Saffran & Kirkham, 2018). An emerging area of statistical learning research investigates learning from an input containing multiple patterns. Multilingual learners represent one instance of this challenge—their input includes two separate languages that must be segregated such that each language can be learned. However, even when we consider a monolingual learner who is only exposed to one dialect of one language, there are myriad linguistic patterns present in that input. In the domain of phonology, an input contains prosodic and segmental cues, within- and between-word phonotactics, and morphologically conditioned phonological patterns such as the patterns for pluralization in English. Although it is clear that infants are able to solve many learning challenges, and even infants learning three or more languages are ultimately successful, much remains to be understood about the process of learning from complex, multidimensional inputs.

The focus of this study is on learning two phonological structures across separate experiments. More specifically, four- and five-year-old children were tasked with learning phonotactic patterns in one experiment and prosodic patterns in the other. Anticipating the results, learning depended on the order in which the experiments were completed. This effect of order is important because it signals that *statistical learning can answer questions about—not only what can be learned—but how episodes of learning interact with each other*. Put slightly differently, our study suggests that statistical learning research that incorporates multiple patterns can be used to explore learning of a pattern even when it is absent from, irrelevant to, or in conflict with the current learning episode.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2021 doi:10.20944/preprints202105.0777.v1

In our review of the literature, we cover several studies with multidimensional inputs. Learners are typically presented with two incompatible or competing inputs. In general, this literature suggests that it is difficult for learners to simultaneously retain knowledge of both inputs. Furthermore, success is often driven by the inputs' phonological properties.

Interactions in Statistical Learning

Weiss, Gerfen, and Mitchel (2009) report four statistical word segmentation experiments exploring how adults interpret and learn from two inputs presented in sequence. In statistical word segmentation tasks, participants hear a continuous stream of syllables like *bətigusɪtʃəvivʊbosætogʊtʃa*. Some syllables always occur in a sequence, such as *bə*, *ti*, and then *gu*, meaning that *bətigu* functions like a word in the stream. However, Weiss et al. interleaved two inputs in the stream such that *bətigu* functioned as a word during some sections of the stream but not others. The participants were tested on their ability to discriminate words like *bətigu* from part-words like *tigusɪ*. Adults learned the patterns from both inputs when each input was spoken by a different talker, but not when the same talker produced both inputs. Weiss et al. suggest that learning fails when participants conflate the statistics of the two inputs, as may be expected when all stimuli come from a single talker.

Gebhart, Aslin, and Newport (2009) conducted a similar statistical word segmentation study with adults. Participants heard just one talker produce both inputs, but the inputs were blocked such that participants heard one in its entirety and then the other. Participants typically only learned the pattern in the first input, although learning of the second pattern was observed when participants were explicitly told to listen for two distinctly patterned inputs, when they heard a pause between each input, as well as when participants heard the second input for three times longer than the first. Gebhart et al. conclude that statistical learning from two inputs shows

 doi:10.20944/preprints202105.0777.v1

a primacy effect—the first of two structurally different streams is likely to be remembered, but not the second. This primacy effect may reflect a learning bias that favors what comes first (see Bulgarelli & Weiss, 2016 for additional discussion), but the authors also propose that, beyond retention, the learning of the first language interferes with learning of the second.

Turning to statistical learning in infants, Benitez, Bulgarelli, Byers-Heinlein, Saffran, and Weiss (2020) observe that 8-month-olds struggle to learn the statistics of the second of two syllable streams. Across several experiments, infants failed to reliably extract words from the second input, even when each input was signaled by a different pitch quality and accent. Although this study does not provide direct evidence for a primacy effect, the first experiment demonstrated that each input was learnable when presented in isolation. Thus, it was the presence of two competing inputs—each with its own phonological patterns—that impeded learning, consistent with Gebhart et al.'s (2009) proposal of a primacy effect.

In contrast to the work with adults, Benitez et al. (2020) observed that, for infants, indexical cues like pitch and accent were insufficient to allow learning of the second of two inputs. A relative weakness of indexical cues was also reported in an infant study by Potter and Lew-Williams (2019). Those authors explored how infants use different types of cues to attune to linguistic structure. They exposed infants to one structured input (either AAB as in *le-le-di* or ABA as in *le-di-le*); that input was embedded in the middle of an unstructured input (16 trisyllables without an internal pattern, such as *foi-nah-vuh*). The authors then varied the cues that signaled the structure—a unique talker, a unique phoneme inventory, or both. When the structured input comprised unique phonemes, infants learned it regardless of whether it was produced by the same talker that produced the unstructured input. Without unique sounds, however, infants were not able to use talker as a cue to learn the structured language. Potter and

 doi:10.20944/preprints202105.0777.v1

Lew-Williams conclude that infants can learn a pattern in the presence of a competing input, but it appears that phonological cues like the phoneme inventory are more informative of the target pattern than indexical cues like talker.

In a study involving similar phonological cues to those examined here, Thiessen and Saffran (2003) observe developmental changes to the phonological cues to which learners attend. Those authors conducted a statistical word segmentation study in which infants were exposed to just one input, for example, where *da* was always followed by *pu*, and *bu* by *go*. However, some infants heard a syllable stream in which stress was consistent with English (weak-strong stress on *DApu* and *BUgo*) whereas other infants heard a stress pattern uncommon in English (strong-weak stress on *daPU* and *buGO*). Nine-month-old infants appeared to ignore statistical cues and instead segment words based entirely on the stress pattern. In contrast, seven-month-old infants used the statistical cues regardless of the stress pattern. Thiessen and Saffran argue that the statistics of syllable order may be an earlier developing phonological cue to word boundaries, but by nine months, word stress is the primary phonological cue for segmenting the speech stream.

Finally, phonology played a surprising role in what infants learned in a study by Gerken and Quam (2017). In this study, 11-month-olds were exposed to just one input. Infants heard novel CVCV words containing a target phonological pattern, either shared place of articulation (*poba* contains two labials) or shared voicing (*dova* contains two voiced consonants). Although only one pattern was present in the exposure words, some infants heard the words in an order that allowed for a local phonological generalization, for example, when two or three adjacent words started with the same consonant. When local generalizations were present, infants did not appear to learn the more global phonological patterns for place of articulation or voicing. When those local generalizations were removed, however, infants learned the more general patterns.

 doi:10.20944/preprints202105.0777.v1

In sum, learning from two incompatible or competing inputs poses a challenge to learners across the lifespan. Although adults are sometimes able to learn patterns across multiple inputs, they often only learn the first pattern presented. Furthermore, both adults and infants rely on phonological cues that signal the pattern in the input. When two patterns are present, phonology may help learners track both, but some phonological cues appear to outweigh others or lead to unintended generalizations. This final point, that learners may apply generalizations unexpectedly, is especially relevant to the present study. Our focus was on the ability of preschool-aged children to learn and apply two distinct phonological patterns to their own speech (prosodic and segmental patterns, similar to Thiessen and Saffran, 2003). Although the patterns were distinct, children completed both experiments, allowing us to examine unintended generalizations across experiments.

Method

Two experiments—one targeting phonotactics and the other prosody—were originally designed to be interpreted separately, and they focus on different dependent measures to track learning. However, individual participants completed both experiments, and the order of experiments was counterbalanced across participants. This counterbalancing allowed us to examine an interaction based on experiment order which is most readily interpretable from the perspective of a single study. Thus, we present both experiments under a single methods section.

Participants

A total of 41 children between the ages of 4 and 5 years (see Table 1) were recruited for the study. Ten children were not included in the analyses because they did not participate for all five days, and they left one or both experiments incomplete. Two additional children were removed

 doi:10.20944/preprints202105.0777.v1

because standardized testing indicated that they had a speech sound disorder. The remaining 29 children (17 females and 12 males) were included in analyses.

All participants met the following criteria for typical development. All children passed a hearing screening of pure tones at 500, 1000, 2000, and 4000 Hz at 20 dB. All children received standardized test scores at or above one standard deviation below the mean (standard scores above 85). Additionally, parents were asked about the child's development, and for the 29 participants included, no concerns were raised.

Normative data were collected across a range of areas: speech production (Goldman-Fristoe Test of Articulation-2; GFTA-2; Goldman & Fristoe, 2000), nonverbal skill (Columbia Mental Maturity Scale; CMMS; Burgemeister, Blum, & Lorge, 1972), receptive vocabulary (Peabody Picture Vocabulary Test-4; PPVT-4; Dunn & Dunn, 2007), expressive vocabulary (Expressive Vocabulary Test; EVT; Williams, 2007), expressive syntax (Structured Photographic Expressive Language Test-3; SPELT-3; Dawson, Stout, & Eyer, 2003), and nonword repetition accuracy (Dollaghan & Campbell, 1998). Table 1 below provides these normative data, as well as the participants' mean age in months, the age range, and average accuracy in the two experiments. Because the critical variable in this study is the experiment order, the normative data are presented separately for the phonotactics first and prosody first groups, and a *t*-test comparison between the groups is presented in the rightmost column. No significant difference, where $p < .05$, was observed. We also note that scores from the CMMS, PPVT-4, the EVT, and the SPELT-3 indicate that this group of participants possessed above-average cognitive and language skills.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2021 doi:10.20944/preprints202105.0777.v1

Table 1. Averages and standard deviations for age in months, standardized test scores, a nonword repetition task, and average production accuracy for each experiment. A statistical comparison of the two experiment order groups appears in the rightmost column.

	M (SD)		*t* statistic (*p* value)*	
	Phonotactics First	Prosody First		
Mean age in months[a]	56.4 (52-67)	58.8 (45-69)	0.96	(*p* = 0.32)
BBTOP standard scores	99.77 (12.47)	102.88 (9.12)	-0.75	(*p* = 0.46)
CMMS standard scores	116.15 (11.51)	111.31 (9.41)	1.2	(*p* = 0.24)
EVT standard scores	109.62 (7.29)	110.88 (9.13)	-0.39	(*p* = 0.70)
PPVT-4 standard scores	119.85 (11.65)	114.31 (9.65)	1.35	(*p* = 0.19)
SPELT-III standard scores	113.85 (7.38)	114.56 (8.22)	-0.24	(*p* = 0.82)
Nonword repetition percent phonemes correct	78.68 (7.76)	74.65 (12.47)	0.94	(*p* = 0.36)
Average accuracy: Phonotactics[b]	5.64 (0.25)	5.43 (0.43)	1.49	(*p* = 0.15)
Average accuracy: Prosody[c]	8.64 (0.39)	8.42 (0.64)	1.05	(*p* = 0.30)

Note. BBTOP = Bankson-Bernthal Test of Phonology, CMMS = Columbia Mental Maturity Scale, EVT = Expressive Vocabulary Test-1, PPVT-4 = Peabody Picture Vocabulary Test-4, SPELT-III = Structured Photographic Expressive Language Test-3. [a]The number in parentheses for ages in months is the range of ages rather than the standard deviation. [b]Accuracy in the phonotactics experiment is on a scale from 0-6. [c]Accuracy in the prosody experiment was averaged across 2-syllable words (scale 0-6) and 4-syllable words (scale 0-12), resulting in a derived scale of 0-9.

*No statistical comparisons of the phonotactics first and prosody first orders were significant.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2021 doi:10.20944/preprints202105.0777.v1

Materials

This study relies on speech production to measure learning. Children under the age of six still produce speech errors and are developing their knowledge of phonology (McLeod & Crowe, 2018). Thus, we ask whether passive, perceptual learning from a familiarization input influences children's production accuracy for test items. The target patterns that children produced were word-medial consonant sequences in the phonotactics experiment, as well as strong-weak and weak-strong stress patterns in the prosody experiment. The familiarization and test materials for both the phonotactics and prosody experiments are presented in Table 2.

Table 2. The learning targets, familiarization items, and test items for the two experiments. Syllable boundaries are indicated with a period. In the high experimental frequency condition, participants heard all three familiarization items. In the low experimental frequency condition, participants only heard the italicized familiarization item.

	Target	Familiarization Items			Test Items
Phonotactics Experiment	/pt/	dap.tən	zeɪp.təs	*sɛp.təf*	bip.təm
	/zm/	koz.mət	lɪz.məs	*taɪz.mək*	pɛz.mef
	/mk/	gum.kəf	tæm.kən	*dɪm.kəs*	fom.kəp
	/fp/	nɪf.pən	ʃeɪf.pək	*kof.pət*	mæf.pəm
Prosody Experiment	2-syllable SW	re.də	ti.də	*do.sə*	po.fə
	4-syllable SW	do.lə.re.sə	so.lə.ti.rə	*la.tə.so.rə*	mi.fə.po.bə
	2-syllable WS	lə.do	tə.re	*sə.la*	bə.mi
	4-syllable WS	də.ti.rə.la	rə.so.tə.do	*tə.la.sə.re*	pə.fa.mə.be

 doi:10.20944/preprints202105.0777.v1

In the phonotactics experiment, the learning targets were word-medial consonant sequences (Munson, 2001). The targets appeared in nonsense words (hereafter referred to as "items") with a CVCCVC shape and stress on the first syllable. All items started with a unique CV sequence and differed from other items by at least three phonemes. Phonotactics familiarization and test items were paired with colorful make-believe animals (Ohala, 1999), and children were told that the nonsense words were the names of the animals. The four target consonant sequences were chosen because consonant sequences are relatively difficult, making it likely that children would sometimes produce them in error, and learning could be measured. We note that data from the phonotactics experiment—when that experiment was completed first—are reported in Richtsmeier and Goffman (2017). All other data have not been reported elsewhere.

In the prosody experiment, the learning targets were prosodic contours, that is, one of two different stress patterns. The first pattern was strong-weak (SW), such as on the noun *REC-ord*; the second pattern was weak-strong (WS) as on the verb *re-CORD*. These patterns appeared in both 2-syllable and 4-syllable items composed of CV syllables, or four total targets. The prosody familiarization and test items were paired with colorful aliens (Gupta et al., 2004). Prosodic contours were chosen because developmental data show that children have not yet reached adult levels of mastery, as indicated by omissions of unstressed syllables as well as acoustic and motor analyses of the WS stress pattern (Ballard, Djaja, Arciuli, James, & van Doorn, 2012; Gladfelter & Goffman, 2013; Goffman, 1999; Goffman, Gerken, & Lucchesi, 2007; Goffman & Malin, 1999).

The phone and biphone frequencies for items from both experiments were calculated using the online Phonotactic Probability Calculator (Vitevitch & Luce, 2004),

 doi:10.20944/preprints202105.0777.v1

https://calculator.ku.edu/phonotactic/about). These frequencies were matched across familiarization and test items. None of the items had phonological neighbors based on a search of the Washington University Speech and Hearing Laboratory's Neighborhood Database (http://128.252.27.56/Neighborhood/NeighborHome.asp). Test items began and ended with labial consonants, and the word-medial sequences in the phonotactics experiment contained at least one labial. The purpose of including many labial consonants was to allow for tracking of lip and jaw movements in kinematic analyses, although those analyses are not reported here.

Recordings of all familiarization and test items were obtained from seven adult female speakers of a Midwestern dialect of American English. The recordings were made in a sound booth following model productions made by the first author. This process was implemented to ensure that acoustic cues for medial consonants and the prosodic contours were produced faithfully. Recordings were later scrubbed of acoustic artifacts and scaled for intensity using Praat software (Boersma & Weenink, 2021). Productions from five of the talkers were used for the familiarization items; productions from the other two talkers were used for the test items.

Experimental Frequency. In both experiments, participants were familiarized with the learning targets during a perceptual familiarization phase, and the experimental frequency of the targets varied as a within-subjects factor, with two targets in the low experimental frequency condition and two in the high experimental frequency condition. Children were familiarized with low experimental frequency targets in just one familiarization item (the items in italics in Table 2). Participants heard that item five times from a single talker. High experimental frequency targets appeared in three familiarization items. Participants heard each item five times, each from a different talker. Thus, high experimental frequency was a combination of high word-type frequency and talker variability (Richtsmeier, Gerken, & Ohala, 2011).

 doi:10.20944/preprints202105.0777.v1

Several previous studies suggest that high experimental frequency can lead to greater production accuracy in children (Edeal & Gildersleeve-Neumann, 2011; Plante, Bahl, Vance, & Gerken, 2011; Richtsmeier, Gerken, Goffman, & Hogan, 2009; Richtsmeier & Moore, 2020), a finding that is consistent with the augmentative effect that high natural language frequency has on production accuracy (Beckman & Edwards, 2000; Edwards, Beckman, & Munson, 2004; Masdottir & Stokes, 2016; Munson, 2001; Storkel, 2015). The assignment of items to the two experimental frequencies was counterbalanced across four lists. The four lists also allowed make-believe animals in the phonotactics experiment and aliens in the prosody experiment to be assigned to different items.

Procedure

The procedures, including informed consent, were approved by the Internal Review Board at Purdue University. Children participated over five weeks, one visit per week. The first session included only testing; participants completed a hearing screening, the SPELT-III language test, and the CMMT nonverbal skill test. Other normative data were collected following the experiment during the other four sessions. The first experiment was completed at the start of the second and third sessions, and the second experiment was completed at the start of the fourth and fifth sessions. Similar numbers of children completed the phonotactics experiment first (n = 16) or the prosody experiment first (n = 13). All sessions were held in a quiet room in a university building. Throughout each session, participants were seated in a Rifton chair with an attachable tabletop, approximately 10 feet from a monitor and speakers. Caregivers were seated nearby.

Before the start of the phonotactics experiment, the experimenter explained that the child would hear the names of “funny, make-believe animals”, and that the child’s task during familiarization was to watch the animals and listen to their names. Before the start of the prosody

 doi:10.20944/preprints202105.0777.v1

experiment, the experimenter explained that the child would hear the names of "aliens from another planet" and that they should watch and listen during the familiarization. Thus, the instructions were similar except for the referents to be learned. The second experiment was nevertheless described as a new experiment and different from what had come before.

The experiments were controlled by Paradigm software (Paradigm, 2015). Each experiment began with familiarization, during which Paradigm presented items in random order. Familiarization was immediately followed by the first test block, and the second test block was completed in the subsequent session a week later. Test items were presented in a predetermined, pseudorandom order, and the same word was repeated no more than twice in a row.

During test blocks, participants were told that they would repeat the names of some new animals or aliens. Children heard and repeated each item immediately. Although children typically required one or two prompts for the first few productions of the first test block, they eventually learned the task and were able to proceed without prompts. Children had nine opportunities to produce each item during a test block. Cases where children did not produce an item were minimal (20 missing productions for the phonotactics experiment; 13 missing productions for the prosody experiment; less than 1% of all attempts), and for all participants, there were always five or more productions of each word in each test block.

Analysis

Children's productions were recorded digitally for transcription and acoustic analysis. The dependent measure of interest for the phonotactics experiment was production accuracy of the word-medial consonant sequence and was based on transcription. Transcriptions were made by the first author and were converted to points based on a system adapted from Edwards et al.

 doi:10.20944/preprints202105.0777.v1

(2004). A correctly produced consonant was given a score of 3. A consonant that differed from the target by one feature (that is, by voicing, place of articulation, or manner of articulation) was given a score of 2. Any other consonant or a consonant sequence was given a score of 1. If no consonant was heard, a score of 0 was given. A second transcriber scored 424 word-medial sequences from 19 of the participants, or approximately 20% of the data. Reliability between the two sets of transcriptions was 90.0% overall (82.3% for the first consonant; 97.6% for the second consonant). Lower accuracy for the first consonant is consistent with studies of similar experimental items (Richtsmeier & Moore, 2020). It likely reflects challenges related to producing and perceiving codas, as the first consonant of the sequence was the coda of the first syllable.

Dependent measures of interest for the prosody study included three ratios of different acoustic markers of stress: ratios of duration, pitch/fundamental frequency, and amplitude (for example, Kehoe, Stoel-Gammon, & Buder, 1995). Omitted or inaudible syllables—based on transcriptions by the first author—comprised a fourth dependent measure. The acoustic measures were analyzed using Praat software (Boersma & Weenink, 2021). The beginning and ending of vowels were first demarcated. Durations were equivalent to the lengths of the demarcated vowel regions. Pitch and intensity were operationalized as the averages across the vowel region. Ratios were then calculated by dividing the value of the first syllable by the value of the second syllable, or σ1/σ2. For four-syllable words, two ratios were collected: σ1/σ2 and σ3/σ4.

Ratios were not calculated for 555 missing productions (26.6%) including productions that children did not make, productions in which one syllable was omitted, or productions that were missing due to experimenter error. From the remaining 1,533 productions, pitch and intensity ratios were removed if the production was whispered, made in creaky voice, contained

 doi:10.20944/preprints202105.0777.v1

acoustic artifacts like foot tapping, or was more than 2 standard deviations from the participant's mean. There were 48 duration ratios (2.1%), 215 pitch ratios (9.4%), and 53 intensity ratios (2.3%) removed for these reasons. Due to missing data, mean pitch ratios were missing for 12 words across 8 participants, and mean intensity ratios were missing for 5 words across 4 participants; all mean duration ratios could be calculated. We also note that the recordings for six participants (all of whom completed the prosody experiment first) contained a high-frequency artifact resembling a square wave. It was created by a sound mixer for unknown reasons. Because the noise started at 1000 Hz, it was not expected to interfere with the measures of pitch and intensity. Data from these six participants were therefore included in the acoustic analyses.

A summary of the inferential statistical analyses is presented in Table 3. The transcription-based points from the phonotactics experiment, as well as the three acoustic ratios and the omitted-syllable counts from the prosody experiment, were entered separately into linear mixed-effects models in R statistical software using the lmerTest package (Kuznetsova, Brockhoff, Christensen, & Jensen, 2020). Mixed effects models are ideal for evaluating incomplete data sets such as the prosody dataset here. We followed recommendations for mixed-model analyses described by Baayen, Davidson, and Bates (2008). In particular, we began with baseline models of main effects that were then compared with more specific models containing interactions. For the baseline models, the main effects of experiment order, experimental frequency, and session were included for both experiments; stress pattern was included as an additional main effect for the prosody experiment. Random effects for participant intercepts were included in all models.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2021 doi:10.20944/preprints202105.0777.v1

Table 3 – A comparison of the dependent variables, independent variables, and statistical models used for the phonotactics and prosody experiments.

	Phonotactics Experiment	Prosody Experiment
Dependent Measures	• Transcription-based accuracy	• Duration Ratios • Pitch Ratios • Intensity Ratios • Omitted Syllables
Independent Variables	• Experiment order • Experimental frequency • Session	• Experiment order • Experimental frequency • Session • Stress pattern
Random Effects	• By-subject Intercepts	• By-subject Intercepts
Simple Model	• Accuracy predicted by experiment order + experimental frequency + session	• Ratios and omitted syllables predicted by experiment order + experimental frequency + session + stress pattern
Alternative Model	• Accuracy predicted by experiment order × experimental frequency + session	• Ratios and omitted syllables predicted by experiment order × experimental frequency + session + stress pattern

 doi:10.20944/preprints202105.0777.v1

To evaluate the presence of interactions between experimental frequency and experiment order, a second model was assessed for each experiment in which experimental frequency and experiment order were allowed to interact. The two models were then compared using a likelihood ratio test that was implemented with the anova function in R (Kuznetsova, Brockhoff, & Christensen, 2017), and the optimal model was interpreted for significant effects.

Results

Phonotactics Experiment

Figure 1 presents the average accuracy in both the high and low experimental frequency conditions across the two experimental orders. In the figure, accuracy is collapsed across sessions, words, and multiple productions. Mean accuracy across all conditions and participants was 5.53. Visual analysis of the figure suggests that participants were slightly more accurate in the high experimental frequency condition when the phonotactics experiment came first (see also Richtsmeier & Goffman, 2017), but when the phonotactics experiment came second, they were more accurate in the low experimental frequency condition.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2021 doi:10.20944/preprints202105.0777.v1

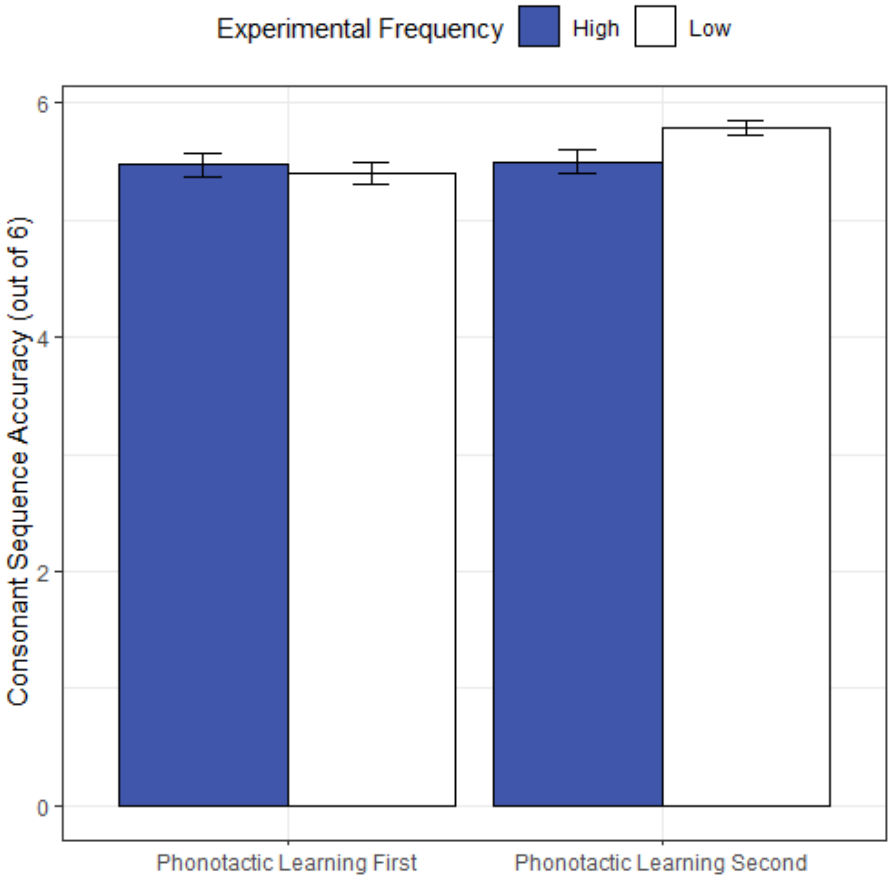

Figure 1 – A comparison of the effect of experimental frequency on consonant sequence accuracy for the two experiment orders. The bars reflect the full range of accuracy values from 0 to 6. Error bars reflect 95% confidence intervals.

The baseline mixed effects model included experiment order, experimental frequency, and session as main effects. Using a log likelihood ratio test, the baseline was then compared with an alternative model in which experiment order and experimental frequency were allowed to interact. The results of the model comparison appear in Table 4 below. The alternative model had lower information criterion scores (AIC and BIC) and a lower log likelihood value. Furthermore, it was significantly better when explaining the data ($\chi^2 = 29.98$, $df = 1$, $p < .001$).

 doi:10.20944/preprints202105.0777.v1

Table 4 – The results of the model comparison for the phonotactics experiment. The baseline model with main effects was compared to an alternative model in which experiment order and experimental frequency were allowed to interact.

Model	df	AIC	BIC	Deviance	χ^2	df	*p*
Baseline model	6	4862.2	4896.0	4850.20	29.98	1	<.001
Alternative model	7	4834.2	4873.7	4820.22			

As the alternative model was significantly better at explaining the data, it is summarized in Table 5 below. Although there was a main effect of experimental frequency (β = .29, *SE* = .05, *t* = 5.79, $p < .001$), the interaction of experimental frequency and experiment order was significant (β = -.37, *SE* = .07, *t* = -5.50, $p < .001$). To better understand that interaction, separate mixed-effects analyses were completed to examine experimental frequency and session in each experiment order condition.

 doi:10.20944/preprints202105.0777.v1

Table 5 – Summary of the alternative mixed effects model of the phonotactics experiment. Statistically significant fixed effects are shown in bold. The number of observations was 2068.

Fixed Effects	*β*	SE	Df	*T*	*p* (>|*t*|)
Intercept	**5.49**	**0.10**	**34.37**	**52.44**	**<.001**
Experiment Order	-0.02	0.14	32.66	-0.18	.656
Experimental Frequency	**0.29**	**0.05**	**2038.99**	**5.79**	**<.001**
Session	0.01	0.03	2039.01	0.45	.862
Experiment Order × Experimental Frequency	**-0.37**	**0.07**	**2039.03**	**-5.50**	**<.001**

Random Effects	Variance	Standard Deviation
Participant (intercept)	0.12	0.35

Accuracy was marginally lower in the low experimental frequency condition for the phonotactic learning first data (β = -.08, *SE* = .05, *t* = -1.66, *p* = .097); accuracy was significantly higher in the low experimental frequency condition for the phonotactic learning second data (β = .29, *SE* = .04, *t* = 6.51, *p* < .001). The results for the phonotactic learning second condition are surprising because high experimental frequency has typically been reported to increase children's production accuracy relative to low experimental frequency (for example, Plante et al., 2011; Richtsmeier et al., 2009; Richtsmeier & Goffman, 2017; Richtsmeier & Good, 2018).

 doi:10.20944/preprints202105.0777.v1

Prosody Experiment

Figure 2 presents a summary of the three acoustic ratios. Mean ratios were calculated by averaging across productions within and across sessions; for the four-syllable words /mifəpoʊbə/ and /pəfɑməbeɪ/, ratios were also averaged across the two syllable groups ([σ1/σ2 + σ3/σ4] ÷ 2). Visual analysis of the ratios suggests robust acoustic contrasts when comparing the SW stress pattern to the WS pattern, but no consistent differences related to experimental frequency.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2021 doi:10.20944/preprints202105.0777.v1

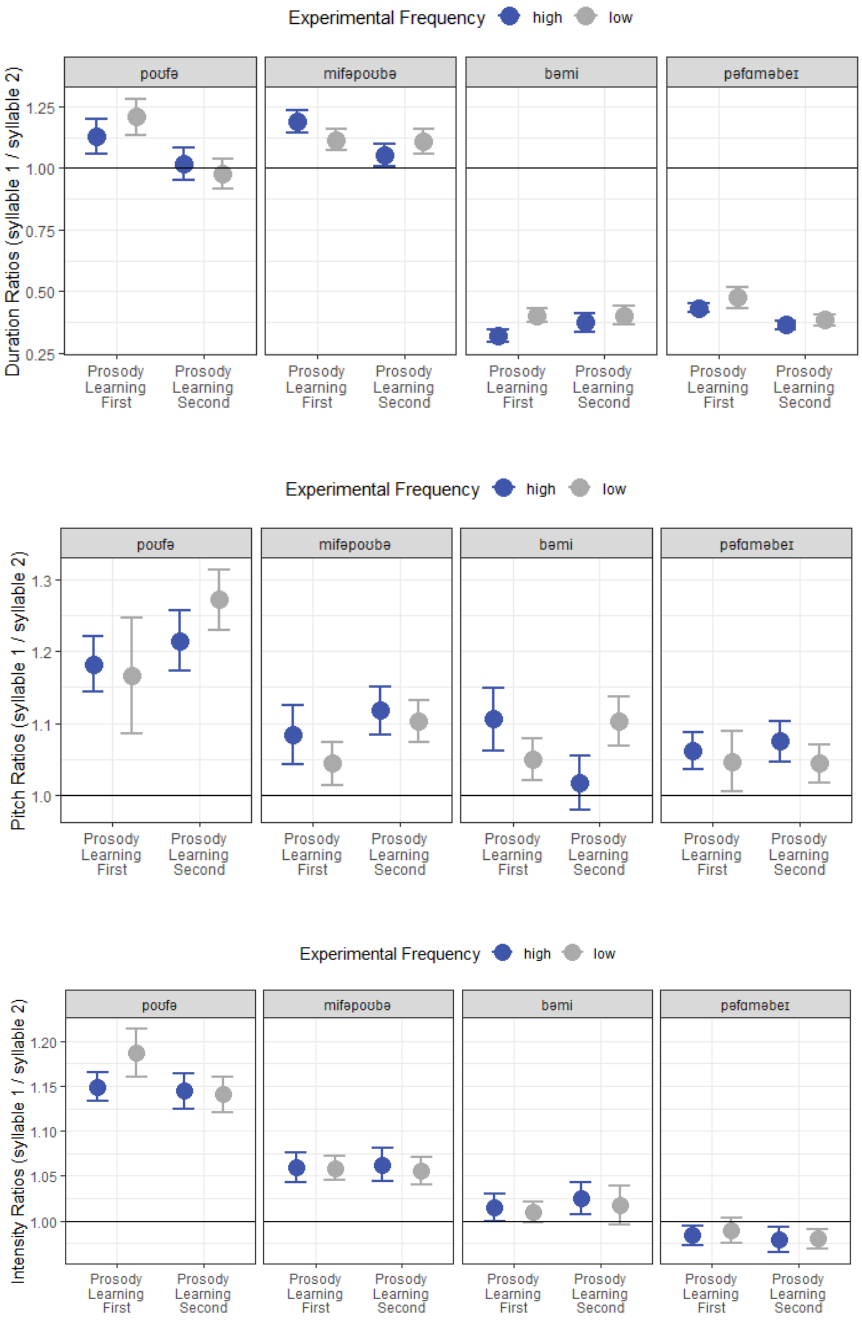

Figure 2 – Duration Ratios (top graph), pitch ratios (middle graph), and intensity ratios (bottom graph). The horizontal black lines in each graph reflects a ratio of 1, or equivalent measurements for the first and second syllable. Ratios greater than 1 generally indicate a SW pattern, and ratios less than 1 indicate a WS pattern.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2021 doi:10.20944/preprints202105.0777.v1

The baseline mixed effects model included experiment order, experimental frequency, session, and stress pattern as main effects. Equivalent to the analysis of the phonotactics experiment, the baseline was then compared with an alternative model in which experiment order and experimental frequency were allowed to interact using a log likelihood ratio test. For duration and intensity ratios, the alternative models did not provide a better fit ($\chi^2_{duration} = 0.25$, $df = 1$, $p = .615$; $\chi^2_{intensity} = 0.004$, $df = 1$, $p = .949$). In the case of pitch ratios, the model with the interaction provided a significantly worse fit compared to the baseline model ($\chi^2_{duration} = 7.55$, $df = 1$, $p = .006$). Thus, in the analyses of three acoustic correlates of stress, experimental frequency did not interact with experiment order. The ANOVAS comparing the baseline and alternative models, as well as the full baseline models of all three ratios, are presented in Appendix A. Here, we present a brief summary of the findings.

There were consistent differences between the SW and WS stress patterns for all three acoustic measures. Regarding durations, participants produced ratios near 1.0 for the SW stress pattern, but ratios less than .5, or second syllables twice as long as first syllables, for the WS pattern ($\beta_{duration} = -0.70$, $p < .001$). Pitch was higher on the first syllable for all items (ratios > 1.0) but highest for the SW pattern ($\beta_{pitch} = -0.06$, $p < .001$). Intensity ratios were greater than 1.0 for the SW pattern, indicating a louder first syllable. Intensity ratios were lower for the WS pattern and averaged below 1.0 for /pəfɑməbeɪ/ ($\beta_{intensity} = -0.10$, $p < .001$). Most importantly, there was no significant effect of experimental frequency in any analysis (all ps > .100). This final result indicates that the acoustic ratio analyses lacked a learning effect attributable to the familiarization and the relative frequencies of the different stress patterns.

The final analysis of the prosody experiment considered inaudible or omitted syllables. Figure 3 presents the number of omitted syllables for each of the four test items. With just 14

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 May 2021 doi:10.20944/preprints202105.0777.v1

omitted syllables observed, the dataset is quite small. Furthermore, one child provided 8 of the 9 omitted syllables for /pəfɑməbeɪ/ in the high experimental frequency, prosody learning second condition. The low number of omitted syllables is to be expected given that the children were all typically developing, and more than half were over the age of four. By that age, syllable omission is quite rare in typically developing children (Roberts, Burchinal, & Footo, 1990).

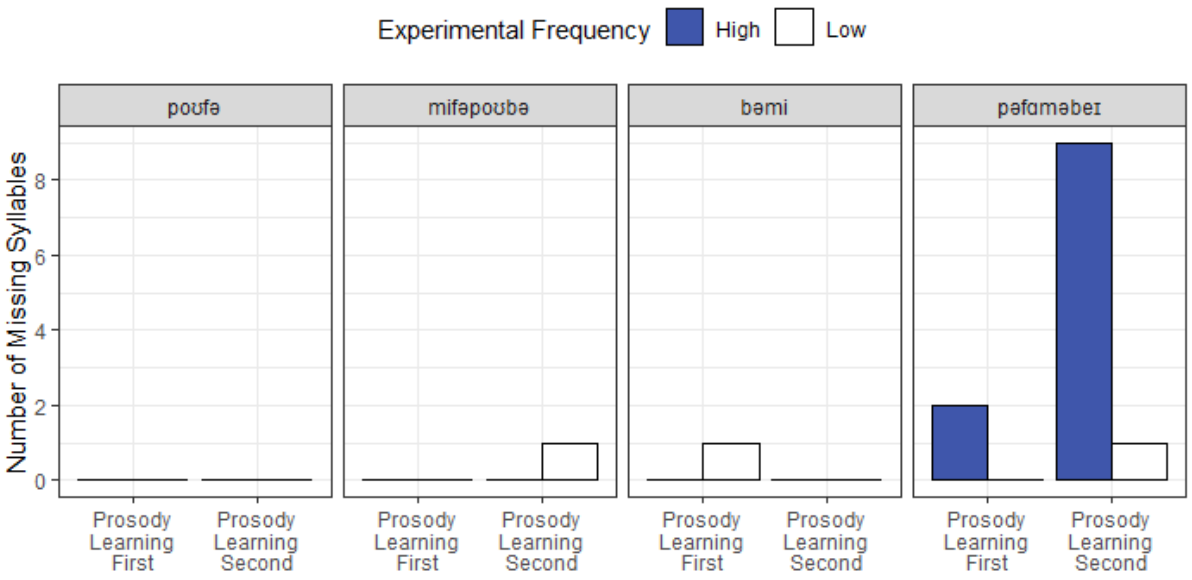

Figure 3 – Omitted syllable counts by word and condition. All omitted syllables were unstressed, specifically, the /fə/ in /mifəpoʊbə/, the /bə/ in /bəmi/, or the /pə/ in /pəfɑməbeɪ/.

Here, we present a summary of the findings from the baseline model because the alternative model was not better at explaining the results ($\chi^2 = 0.50$, $df = 1$, $p = .481$). There was a trend towards a greater number of omitted syllables in words with the WS stress pattern ($\beta = -.80$, $SE = .43$, $t = -1.85$, $p = .073$); all other effects were not significant ($p > .3$). The ANOVA comparison of the two models, as well as the full baseline model, are included in Appendix B.

 doi:10.20944/preprints202105.0777.v1

To summarize the results of the prosody experiment, there were consistent differences related to the SW and WS stress patterns, but there was no significant interaction between experiment order and experimental frequency. More generally, with no observed effect of experimental frequency, there was limited evidence for learning of the stress patterns.

General Discussion

In this study, children completed two statistical learning experiments—one focused on learning phonotactics in the form of word-medial consonant sequences, the other focused on learning prosodic modulation in the form of SW and WS stress patterns. Learning was probed by a comparison of high and low experimental frequency conditions. In the high experimental frequency condition, learning targets appeared in multiple items produced by multiple talkers; in the low experimental frequency condition, targets appeared in a single item produced by one talker. A relative difference in the high and low experimental frequency conditions was taken to signal learning. The key finding is that learning of the phonotactic sequences was influenced by the order in which participants completed the two experiments. When participants completed phonotactic learning first, there was a trend towards greater accuracy following high experimental exposure. In contrast, when participants completed phonotactic learning second, participants were significantly more accurate following the low experimental frequency exposure. Therefore, the effect of experimental frequency, and by extension learning, varied depending on the order in which the two experiments were completed. An experiment order by experimental frequency interaction was not observed in the prosody experiment, nor was there a main effect of experimental frequency. Below we describe several reasons why the effect may have been limited to the phonotactics experiment.

 doi:10.20944/preprints202105.0777.v1

The interaction of experiment order and experimental input frequency is consistent with the literature reviewed in the Introduction. Benitez et al. (2020), Gebhart et al. (2009), Potter and Lew-Williams (2019), and Weiss et al. (2009) all found that conflicting language inputs were challenging for participants. In the adult studies, participants often only learned the first of two artificial languages presented in succession. The infant study by Benitez et al. was consistent with the same "primacy effect". The present study may also reflect a primacy effect. When participants completed the phonotactics experiment first, the effect of experimental frequency was consistent with previous findings. It was only when the experiment was completed second that a surprising finding arose.

The results are also consistent with various phonological cues as central to the learnability—or difficulty—of a multidimensional input. When searching for word boundaries, Thiessen and Saffran (2003) find that 7-month-old infants rely on statistical cues, but 9-month-olds ignore the statistics and instead rely on prosodic cues. Gerken and Quam (2017) report that infants are sometimes misled by narrow phonological generalizations, such as repeated word-initial consonants. Potter and Lew-Williams' (2019) infants demonstrate that learning of a structured input—surrounded by unstructured input—is possible when the structure is signaled by a unique inventory of phonemes. Here, the phonotactic and prosody experiments targeted different phonological structures (Kenstowicz & Kisseberth, 2014). Phonotactic generalizations often occur at a segmental level. Prosodic generalizations, in contrast, occur at a metrical or intonational level that spans multiple syllables. Additionally, a variety of cues were given to participants so that they might treat the experiments as separate. These cues included time (a week between experiments), visual referents (make-believe animals for phonotactics and aliens for prosody), and instructions (participants were told that the second experiment was, in fact, a

 doi:10.20944/preprints202105.0777.v1

different experiment). Given these factors, it would appear that the two experiments were relatively well distinguished, at least in terms of the phonological aspects to attend to.

The list of differences above notwithstanding, the interaction of experiment order and experimental frequency informs us that participants did not treat the experiments as separate. Given the clear phonological distinction between targets, the interference is most readily attributable to the phonological learning environment. In particular, participants began with an exposure phase in which they listened to a structured input of nonwords with high and low experimental frequency conditions. Future research is needed to establish a unified account of the various types of phonological interference or cue interaction observed by Gerken and Quam (2017), Potter and Lew-Williams (2018), and here. Furthermore, an account of phonological interference should also account for the type of segmental and prosodic cue integration studied by Thiessen and Saffran (2003), as well as the successful input segregation observed in adults by Gebhart, Weiss, and colleagues. Robust segregation of different statistical learning inputs may not occur until adolescence or adulthood. Of course, in the real world, infants and children are exposed to a vast array of inputs reflecting different rules and patterns, so future research is also needed to better understand the conditions under which even the youngest learners can learn from and segregate a multidimensional input.

Finally, our study is consistent with proposals by Gebhart et al. (2009) and Bulgarelli and Weiss (2016) that the primacy effect likely reflects a kind of interference across experiments. This interference was signaled by a qualitative difference in the direction of the experimental frequency effect that was determined by experiment order. Despite a surface connection to previous findings of interference, the accuracy advantage for low experimental frequency sequences in the phonotactic learning second experiment is remarkable, and to our knowledge, it

 doi:10.20944/preprints202105.0777.v1

is unprecedented. Consider the logic put forth by Richtsmeier et al. (2011) to explain the benefits of a high experimental frequency: They argue that this benefit is consistent with the high frequency advantage seen across language development (Ambridge, Kidd, Rowland, & Theakston, 2015). In fact, Richtsmeier et al. interpret experimental frequency as a simulation of the production advantage for high English frequency consonant sequences, such as when children produce novel words with high-frequency sequences like /blɪk/ more accurately than words with low-frequency sequences like /sfɪk/. These effects have been reported by Edwards et al. (2004), Masdottir and Stokes (2016), Munson (2001), Richtsmeier et al. (2009), Zamuner, Gerken, and Hammond (2004), and many others. To have obtained the opposite of this well-established result is striking.

As such, we consider the experiment order by experimental frequency interaction to be most consistent with the kind of unintended generalization observed by Gerken and Quam (2016). In other words, a relative advantage for low experimental frequency sequences in the phonotactic learning second condition may be a case of the wrong generalization being applied. This is in part because participants in the phonotactic learning second condition probably did not better learn the low experimental frequency sequences. In some sense, such an explanation defies the basic notion of learning, which is closely tied to stimulus frequency (Ambridge et al., 2015). Rather, we argue that high experimental frequency had a kind of damping effect because it was inconsistent with high experimental frequency from the previous experiment. That is, it was inconsistent with the high frequency, prosody-focused items from the prosody experiment.

Additional research is necessary to verify that participants were attempting to impose patterns from the first input onto the second input. Here and in previous studies, the authors have verified that participants did not exhibit an expected pattern, but they have not specifically

 doi:10.20944/preprints202105.0777.v1

probed for the unintended generalization. In other words, what is needed is explicit evidence that participants are applying a pattern from the first experiment to the second experiment. Such a study is warranted, particularly in the area of child speech development.

Children are known to develop patterns in their speech that never appear in the input. Phonological processes such as fronting (underlying /k, g/ are produced as [t, d]; [tæt] for *cat*), stopping (underlying /s, z/ are produced as [t, d]; [tʌn] for *sun*), and gliding (underlying /r, l/ are produced as [w]; [wɑk] for *rock*) are all unexpected patterns in that they are never observed in the child's input. That is, adults do not provide models of phonological processes like fronting, stopping, or gliding. How do children learn these unobserved phonological patterns? One possibility is that they start out as unintended generalizations from statistical learning. For example, a child who hears and imitates several words in a row that begin with initial alveolar stops (*toy*, *tooth*, *dad*, and *doll*) may draw the generalization that word-initial stops are alveolars. If they are later exposed to the word *car*, that generalization could result in a production like [tɑr]. This type of generalization is the focus of ongoing experiments in the first author's lab. Furthermore, using the statistical learning paradigm to study phonological processes may shed light on why phonological processes like fronting are relatively common, including in typical development, whereas processes like backing (underlying /t, d/ are produced as [k, g]; [hæk] for *hat*) are rare.

A notable limitation of the present study is that no learning effects were observed in the prosody experiment. More specifically, there was not a significant main effect in any of the acoustic ratios or in the number of omitted syllables. As such, it may not have been possible to observe experiment order by experimental frequency interactions to either reinforce or limit the interpretation of the phonotactics study.

 doi:10.20944/preprints202105.0777.v1

It may be that our participants were proficient enough when producing strong-weak and weak-strong contours that the results reflect an aspect of phonology that is less amenable to learning. There is some support for such a conclusion. For example, Pollock, Brammer, and Hageman (1993) found that 3- and 4-year-olds were able to consistently use duration, pitch, and amplitude to distinguish polysyllabic nonwords with strong-weak and weak-strong prosody. In a larger study with children up to seven years of age, Ballard et al. (2012) found that children acquiring English were adept at using duration, pitch, and amplitude to signal strong-weak patterns as young as age three. However, the strong-weak patterns produced by seven-year-olds did not reach adult levels of contrast. Our data are consistent with Ballard et al.'s protracted developmental trajectory. In statistical comparisons with the acoustic ratios of the adult model productions, children differed from the adult norms, particularly in the use of pitch for strong-weak patterns (see Appendix C). A similar delay for the strong-weak pattern was observed in kinematic analyses of articulatory stability made by Goffman and Malin (1999). Thus, there was room for learning to be observed in some of the acoustic parameters of prosody. However, relative to the impact of phonemic substitution errors common in the phonotactics experiment, there may have been fewer perceptual consequences to falling short of adult-like prosodic targets because the basic targets of SW and WS were being achieved. In this more nebulous learning space, children may have had fewer incentives to improve their production targets for the prosody items. Regardless of the adequacy of this explanation, further research is needed to better understand the learnability of various phonological targets within the statistical learning paradigm and as applied to child speech development.

In conclusion, the results of this study reflect an interesting case of multidimensional statistical learning. Compared to many previous studies in this area, our study did not include

 doi:10.20944/preprints202105.0777.v1

stimuli that were inherently in conflict. Phonotactics and prosody are different enough that it would be reasonable to expect participants to learn them separately. Nevertheless, learners did not treat the two experiments as separate. When participants completed the phonotactic experiment first, learning was consistent with previous findings, perhaps reflecting a primacy effect for initial learning. When participants completed the phonotactic experiment second, participant accuracy was unexpectedly low for the high experimental frequency condition, indicating interference from the previously completed prosody experiment. Exactly what this interference looks like, and whether it reflects overgeneralization of the patterns from the first experiment, remains to be determined by future studies.

Acknowledgements

This research was supported by grants from the National Institute on Deafness and Other Communication Disorders; R03DC011898 to Peter T. Richtsmeier and R01DC016813 to Lisa Goffman. The authors thank Erica Berlin, Janna Berlin, Antigone Fleck, Shie Kantor, and Meredith Saletta for their assistance with various aspects of data collection and processing.

References

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language, 42*(2), 239-273. doi:https://doi.org/10.1017/S030500091400049X

Baayen, R. H., Davidson, D. J., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory & Language, 59*, 390-412. doi:https://doi.org/10.1016/j.jml.2007.12.005

Ballard, K. J., Djaja, D., Arciuli, J., James, D. G. H., & van Doorn, J. (2012). Developmental Trajectory for Production of Prosody: Lexical Stress Contrastivity in Children Ages 3 to 7 Years and in Adults. *Journal of Speech, Language, and Hearing Research, 55*(6), 1822-1835. doi:https://doi.org/10.1044/1092-4388(2012/11-0257)

Beckman, M. E., & Edwards, J. (2000). Lexical frequency effects on young children's imitative productions. In M. Broe & J. B. Pierrehumbert (Eds.), *Papers in laboratory phonology V* (pp. 207-217). Cambridge, UK: Cambridge University Press.

Benitez, V. L., Bulgarelli, F., Byers-Heinlein, K., Saffran, J. R., & Weiss, D. J. (2020). Statistical learning of multiple speech streams: A challenge for monolingual infants. *Developmental Science, 23*(2), e12896. doi:https://doi.org/10.1111/desc.12896

Boersma, P., & Weenink, D. (2021). Praat: Doing phonetics by computer. Retrieved from http://www.praat.org/

Bulgarelli, F., & Weiss, D. J. (2016). Anchors aweigh: The impact of overlearning on entrenchment effects in statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(10), 1621.

Burgemeister, B. B., Blum, L. H., & Lorge, I. (1972). Columbia mental maturity scale (CMMS)[assessment instrument]. In. New York, NY: Harcourt Brace Jovanovich.

Dawson, J. I., Stout, C. E., & Eyer, J. A. (2003). Structured Photographic Expressive Language Test–Third Edition (SPELT-3)[assessment instrument]. In. DeKalb, IL: Janelle Publications.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research, 41*(5), 1136-1146. doi:https://doi.org/10.1044/jslhr.4105.1136

Dunn, L. M., & Dunn, D. M. (2007). Peabody picture vocabulary test: (PPVT-4)[assessment instrument]. In. Minneapolis, MN: Pearson Assessments.

Edeal, D. M., & Gildersleeve-Neumann, C. E. (2011). The importance of production frequency in therapy for childhood apraxia of speech. *American Journal of Speech-Language Pathology, 20*(2), 95-110. doi:https://doi.org/10.1044/1058-0360(2011/09-0005)

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research, 47*(2), 421-436. doi:https://doi.org/10.1044/1092-4388(2004/034)

Gebhart, A. L., Aslin, R. N., & Newport, E. L. (2009). Changing structures in midstream: Learning along the statistical garden path. *Cognitive Science, 33*(6), 1087-1116.

Gerken, L., & Quam, C. (2017). Infant learning is influenced by local spurious generalizations. *Developmental Science, 20*(3). doi:https://doi.org/10.1111/desc.12410

Gladfelter, A., & Goffman, L. (2013). The influence of prosodic stress patterns and semantic depth on novel word learning in typically developing children. *Language Learning and Development, 9*(2), 151-174.

Goffman, L. (1999). Prosodic influences on speech production in children with specific language impairment and speech deficits: kinematic, acoustic, and transcription evidence. *Journal of Speech, Language, and Hearing Research, 42*(6), 1499-1517.

Goffman, L., Gerken, L., & Lucchesi, J. (2007). Relations Between Segmental and Motor Variability in Prosodically Complex Nonword Sequences. *Journal of Speech, Language, and Hearing Research, 50*(2), 444-458. doi:doi:10.1044/1092-4388(2007/031)

Goffman, L., & Malin, C. (1999). Metrical effects on speech movements in children and adults. *Journal of Speech, Language, and Hearing Research, 42*(4), 1003-1015.

Goldman, R., & Fristoe, M. (2000). *Goldman-Fristoe Test of Articulation-2*. San Antonio, TX: Pearson.

Gupta, P., Lipinski, J., Abbs, B., Lin, P.-H., Aktunc, E., Ludden, D., . . . Newman, R. (2004). Space aliens and nonwords: Stimuli for investigating the learning of novel word-meaning pairs. *Behavior Research Methods, Instruments, & Computers, 36*(4), 599-603. doi:https://doi.org/10.3758/BF03206540

Kehoe, M., Stoel-Gammon, C., & Buder, E. H. (1995). Acoustic correlates of stress in young children's speech. *Journal of Speech and Hearing Research, 38*(2), 338-350.

Kenstowicz, M., & Kisseberth, C. (2014). *Generative phonology: Description and theory*. United Kingdom: Elsevier Science.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1-26. doi:http://dx.doi.org/10.18637/jss.v082.i13

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B., & Jensen, S. P. (2020). lmerTest (Version 3.1-3) [Computer software]. Retrieved from https://github.com/runehaubo/lmerTestR

Masdottir, T., & Stokes, S. F. (2016). Influence of consonant frequency on Icelandic-speaking children's speech acquisition. *International Journal of Speech-Language Pathology, 18*(2), 111-121. doi:https://doi.org/10.3109/17549507.2015.1060525

McLeod, S., & Crowe, K. (2018). Children's consonant acquisition in 27 languages: A cross-linguistic review. *American Journal of Speech-Language Pathology, 27*(4), 1546-1571. doi:https://doi.org/10.1044/2018_AJSLP-17-0100

Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research, 44*(4), 778-792.

Ohala, D. K. (1999). The influence of sonority on children's cluster reductions. *Journal of Communication Disorders, 32*(6), 397-421; quiz 421-392. doi:https://doi.org/10.1016/S0021-9924(99)00018-0

Paradigm. (2015). Paradigm Experimental Software (Version 2.4) [Computer software]. Lawrence, KS: Perception Research Systems Incorporated. Retrieved from http://www.paradigmexperiments.com/

Plante, E., Bahl, M., Vance, R., & Gerken, L. (2011). Beyond phonotactic frequency: presentation frequency effects word productions in specific language impairment.

Journal of Communication Disorders, 44(1), 91-102. doi:https://doi.org/10.1016/j.jcomdis.2010.07.005

Pollock, K. E., Brammer, D. M., & Hageman, C. F. (1993). An acoustic analysis of young children's productions of word stress. *Journal of Phonetics, 21*(3), 183-203. doi:https://doi.org/10.1016/S0095-4470(19)31332-4

Potter, C. E., & Lew-Williams, C. (2019). Infants' selective use of reliable cues in multidimensional language input. *Developmental Psychology, 55*(1), 1-8. doi:https://doi.org/10.1037/dev0000610

Richtsmeier, P. T., Gerken, L., Goffman, L., & Hogan, T. (2009). Statistical frequency in perception affects children's lexical production. *Cognition, 111*(3), 372-377. doi:https://doi.org/10.1016/j.cognition.2009.02.009

Richtsmeier, P. T., Gerken, L., & Ohala, D. K. (2011). Contributions of phonetic token variability and word-type frequency to phonological representations. *Journal of Child Language, 38*(5), 951-978. doi:https://doi.org/10.1017/S0305000910000371

Richtsmeier, P. T., & Goffman, L. (2017). Perceptual statistical learning over one week in child speech production. *Journal of Communication Disorders, 68*, 70-80. doi:https://doi.org/10.1016/j.jcomdis.2017.06.004

Richtsmeier, P. T., & Good, A. K. (2018). Frequencies in perception and production differentially affect child speech. *Journal of Speech, Language, and Hearing Research, 61*(12), 2854-2868. doi:https://doi.org/10.1044/2018_JSLHR-S-17-0391

Richtsmeier, P. T., & Moore, M. (2020). Order effects in the perception and production of new words [Preprint]. *Preprints*. doi:https://doi.org/10.20944/preprints202005.0383.v1

Roberts, J. E., Burchinal, M., & Footo, M. M. (1990). Phonological process decline from 212 to 8 years. *Journal of Communication Disorders, 23*(3), 205-217. doi:https://doi.org/10.1016/0021-9924(90)90023-R

Saffran, J. R., & Kirkham, N. Z. (2018). Infant Statistical Learning. *Annual Review of Psychology, 69*(1), 181-203. doi:https://doi.org/10.1146/annurev-psych-122216-011805

Storkel, H. L. (2015). Learning from input and memory evolution: Points of vulnerability on a pathway to mastery in word learning. *International Journal of Speech-Language Pathology, 17*(1), 1-12.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39*(4), 706-716. doi:10.1037/0012-1649.39.4.706

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers, 36*(3), 481-487. doi:https://doi.org/10.3758/BF03195594

Weiss, D. J., Gerfen, C., & Mitchel, A. D. (2009). Speech segmentation in a simulated bilingual environment: A challenge for statistical learning? *Language Learning and Development, 5*(1), 30-49.

Williams, K. T. (2007). *Expressive Vocabulary Test [assessment instrument]* (Second Edition ed.). Minneapolis, MN: Pearson Assessments.

Zamuner, T. S., Gerken, L., & Hammond, M. (2004). Phonotactic probabilities in young children's speech production. *Journal of Child Language, 31*(3), 515-536.