Phonetic effects in child and adult word segmentation

Corresponding author:
Jonah Katz
Dept. of World Languages, Literatures, & Linguistics
West Virginia University
Morgantown, WV, 26501
USA
+1 (304) 293-5121

Michelle W. Moore
Department of Communication Sciences and Disorders
West Virginia University
Morgantown, West Virginia

ABSTRACT

Purpose: To investigate the effects of specific acoustic patterns on word learning and segmentation in 8- to 11-year old children and in college students.

Method: Twenty-two children (ages 8;2 – 11;4) and 36 college students listened to synthesized 'utterances' in artificial languages consisting of 6 iterated 'words', which followed either a phonetically natural lenition-fortition pattern or an unnatural (cross-linguistically unattested)

anti-lenition pattern. A two-alternative forced choice task tested whether they could discriminate between occurring and non-occurring sequences. Participants were exposed to both languages, counterbalanced for order across subjects, in sessions spaced at least 1 month apart.

Results: Children showed little evidence for learning in either the phonetically natural or unnatural condition, nor evidence of differences in learning across the two conditions. Adults showed the predicted (and previously attested) interaction between learning and phonetic condition: the phonetically natural language was learned better. The adults also showed a strong effect of session: subjects performed much worse during the second session than the first.

Conclusion: School-age children not only failed to demonstrate the phonetic asymmetry demonstrated by adults in previous studies, but failed to show strong evidence for any learning at all. The fact that the phonetic asymmetry (and general learning effect) was replicated with adults suggests that the child result is not due to inadequate stimuli or procedures. The strong carryover effect for adults also suggests that they retain knowledge about the sound patterns of an artificial language for over a month, longer than has been reported in laboratory studies of purely phonetic/phonological learning.

INTRODUCTION

The question of how domain-general statistical learning abilities interact with domain-specific aspects of speech perception is central in the linguistic and cognitive sciences (e.g. Frost, Monaghan, & Tatsumi, 2017; Johnson & Jusczyk, 2001; Johnson & Seidl, 2009; Saffran, Newport, & Aslin, 1996a). Statistical learning also has been studied in communication sciences

and disorders as a way to compare and contrast the implicit learning abilities of people with communication disorders versus their typical peers (e.g. Evans, Saffran, & Robe-Torres, 2009; Plante, Gomez, & Gerken, 2002). While conducting studies of this nature is critical for understanding the language learning process across development and how to maximize learning opportunities, there are many open questions about the extent to which the properties of the task and characteristics of the participants affect performance outcomes. This study focuses on effects of specific acoustic properties and long-term memory in statistical learning using a word-segmentation paradigm.

Statistical learning and the word-segmentation paradigm

While there is no generally agreed-upon set of boundary conditions for what constitutes 'statistical learning', we follow Frost, Armstrong, & Christiansen (2019) in using the term generally to refer to perceiving and learning temporal and spatial patterns in the environment. While it is quite clear that this general type of ability exists across domains and modalities, we are not committed to the idea that, for instance, linguistic and visual statistical learning are 'the same' (see Siegelman, Bogaerts, Elazar, Arciuli, & Frost, 2018 for discussion and arguments against full domain-generality).

While statistical learning has received an enormous amount of attention in cognitive science (see Frost, et al. 2019 for a systematic review), it is not the only type of learning active in speech and language, and in fact interacts in intricate ways with other types of learning (Romberg & Saffran, 2010). For instance, various types of perceptual learning (Goldstone, 1998) can either form inputs to statistical learning, can be applied to outputs, or can be seen as a form of statistical learning themselves. At a first pass, tracking the recurrence of elements in temporally complex stimuli requires that those elements be grouped into equivalence classes, and

perceptual learning is an important driver of category formation. At the same time, in the domain of linguistic sound patterns studied here, it has been frequently argued that statistical learning is used to optimize sound patterns for the perceptual learning process known as 'attentional weighting' (e.g. Cohen Priva, 2017; Cohen Priva & Gleason, 2020; Harris, 2003; Katz, 2016; Katz & Pitzanti, 2019). Furthermore, the perceptual learning process known as 'unitization' is itself a form of statistical learning. Unitization is essentially the idea that elements which frequently co-occur can eventually be assigned status as unitary features, and it plays an important role in speech perception (Diehl, Lotto, & Holt, 2004). The current study examines how domain-specific aspects of speech perception interact with general statistical learning.

One common way of probing statistical learning abilities is the word-segmentation paradigm, which has been successfully employed with infants (Saffran, Aslin, & Newport, 1996b), school-aged children (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997), and adults (Saffran *et al*., 1996a). In the word-segmentation paradigm, subjects are first exposed to a sequence of syllables in the training phase. Rather than random sequences, the syllables are drawn from a small set of made-up 'words' consisting of a number of syllables, such that syllables that follow each other within a word will always appear in that order, while syllables at the beginning or ending of a word may be preceded or followed by a variety of different syllables depending on the adjacent word. The set of strings consisting of such words is referred to as the 'artificial language' created by the researchers, though of course it lacks many elements of actual real-world human languages. Consider, for instance, the language consisting only of the recurring elements [bapito], [kisofe], and [mufali]. An 'utterance' in this language might be organized as in (1), where '.' denotes a word boundary:

(1) [bapito.kisofe.mufali.kisofe.bapito.mufali.bapito]

In this language, the syllable [ba] is always followed by [pi], and the same is true, *mutatis mutandis*, for all other consecutive syllables within words. That is, given the syllable [ba], there is a 100% chance (the conditional probability) that the next syllable will be [pi]. Given the syllable [to], however, there are three possibilities for the following syllable: it may be any word-initial syllable in the language. If words are concatenated randomly, then the conditional probabilities of [ba], [ki], and [mu] following [to] will each approximate 1/3.

As first pointed out by Harris (1955) and Chomsky (1955), a language learner could use this general type of asymmetry between unit-internal transitions (highly probable) and cross-unit transitions (less probable in the general case) to infer which sequences in a continuous speech stream are more likely to be coherent units. The word-segmentation paradigm suggests that learners of all ages do just this: following the training phase, listeners are tested for their ability to discriminate words in the artificial language from either non-occurring sequences or sequences that occur but span word boundaries (and thus have internal sequences with conditional probability less than 1). For infants, preferential looking is generally used for the test phase; for older listeners, two-alternative forced choice tasks are common ('Which of these is a word in the language you heard?'). A large body of studies find above-chance performance across different kinds of listeners, language designs, experimental procedures, acoustic materials, and conditions of variability (e.g. Evans *et al.*, 2009; Frank, Goldwater, Griffiths, & Tenenbaum, 2010; Karuza *et al.*, 2015; Kim, 2004; Saffran *et al.*, 1996a, 1996b, 1997). While these experiments generally are labeled 'word segmentation', nothing in the experiments nor the theoretical account is

115 specific to the *word* level of constituency: the logic holds equally for segments, syllables,
116 morphemes, prosodic groups, etc.
117 The stimuli in the word-segmentation experiment used here are tightly controlled to
118 eliminate all acoustic differences beyond the lenis and fortis nature – i.e. the relatively greater or
119 lesser degree of energy, respectively – of the consonants involved, which is the target
120 manipulation. As such, these stimuli depart in many ways from natural speech and one could
121 question their relevance to real-world language learning and processing. Previous research,
122 however, gives reasons for cautious optimism that statistical-learning tasks, even those using
123 simplified and artificial stimuli, capture meaningful variation between individuals associated
124 with real-world language outcomes (see Siegelman, Bogaerts, & Frost, 2017, for review of such
125 evidence as well as a number of problematic aspects of such literature). For instance, children's
126 performance on a word-segmentation task similar to the one used here tracks vocabulary
127 knowledge and phonological processing (Spencer, Kaschuk, Jones, & Lonigan, 2015) as well as
128 lexical-phonological ability (Mainela-Arnold & Evans, 2014). Word-segmentation is impaired in
129 children with specific language impairment (Evans et al., 2009), and these results also hold for
130 related statistical-learning tasks such as artificial grammar learning (Lukács & Keméney, 2014)
131 and phonotactic learning (Mayor-Dubois, Zesiger, Van der Linden, & Roulet-Perez, 2014). A
132 number of studies with adults and children suggest that various other statistical-learning related
133 tasks, in both the visual and auditory domain, track individual differences in sentence processing,
134 speech perception, and reading ability (Arciuli & Simpson, 2012; Conway, Karpicke, & Pisoni,
135 2007; McCauley, Isbilen, & Christiansen, 2017; Misyak & Christiansen, 2011). So while the
136 relevance of statistical learning to language seems most obvious for infants who are acquiring a

lexicon, it is apparent that these abilities remain linked to speech and language outcomes throughout the lifespan.

Sound patterns and word-segmentation

In most of the studies mentioned above, researchers were interested in testing the idea that listeners can learn from statistical regularities alone, in the absence of reinforcing cues to constituency. For this reason, most of the stimuli in these experiments were synthesized, keeping duration constant across syllables and using flat contours for intensity and pitch. This guarantees that if subjects show learning, it must be from patterns of phonemes or syllables alone, and not from any acoustic properties used to mark linguistic constituency. In real-world language, of course, there are various phonetic and phonological properties used to mark constituent edges at all levels, from syllables (Browman & Goldstein 1990) to morphemes (Sugahara & Turk, 2009), prosodic words (Selkirk, 1995; Turk & Shattuck-Hufnagel, 2000), and phrases (Pierrehumbert, 1980; Nespor & Vogel, 1986).

Sound patterns associated with prosodic boundaries are important because statistically-based segmentation effects are modulated by prosodic and sub-phonemic acoustic properties. Saffran *et al*. (1996a), for instance, show that English-speaking adults perform better on the word-segmentation task when statistical regularities are reinforced by lengthening the final syllable of every other word. This acoustic pattern roughly matches the English phenomenon of final lengthening, which lengthens the final syllable of prosodic phrases (Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992); more generally, final lengthening is a prosodic feature of many languages (see Gordon & Munro, 2007, for review). Subsequent studies have shown that

159 the reinforcing effect of final lengthening on word-segmentation is present for adult speakers of
160 many different languages (Frost *et al.*, 2017; Tyler & Cutler, 2009).

161 Other studies have revealed that language-specific acoustic patterns facilitate word-
162 segmentation for adult and infant speakers of a number of languages. Initial stress and decreased
163 consonant-vowel coarticulation at word boundaries both aid English-learning 8-month-olds
164 (Johnson & Jusczyk 2001). The stress effect has been replicated for 9-month-olds but not 7-
165 month-olds (Thiessen & Saffran 2003), and can be overridden by either native-language
166 experience (Polka & Sundara 2003) or experimental training immediately before test (Thiessen
167 & Saffran 2007). Final lengthening and language-specific intonational cues aid French-speaking
168 adults (Bagou, Fougeron, & Frauenfelder, 2002; Tyler & Cutler, 2009). Language-specific
169 intonational cues also aid adult speakers of Korean (Kim, 2004), English, and Dutch (Tyler &
170 Cutler, 2009).

171 It is clear from these studies that language-specific sound patterns affect the ease of word
172 segmentation, and that experience with different languages can thus affect which sound patterns
173 are relevant in this regard (e.g. Onnis & Thiessen, 2013; Potter, Wang, & Saffran, 2017;
174 Siegelman, Bogaerts, Elazar, Arciuli, & Frost, 2018). Much less is known, however, about sound
175 patterns that *don't* vary between languages. In particular, some phonetic and phonological
176 patterns are quite widely attested across languages and language families, show limited
177 variability with regard to the prosodic contexts in which they occur, and are claimed to be due to
178 general properties of speech (e.g. Ohala, 1975), audition (e.g. Steriade, 2001) and/or
179 communicative efficiency (e.g. Lindblom 1983). If such explanations are correct, then these
180 relatively invariant patterns may affect word segmentation even by listeners who *lack* native-
181 language experience with such patterns (Katz & Fricke, 2018).

The particular sound pattern we investigate in this study is referred to as *spirantization* in the phonological literature. It is a type of lenition-fortition, or strengthening-weakening, process. It is widely attested across dozens of unrelated languages and tends to occur in specific phonetic contexts across all of those languages; in between two vowels is the most common (Kirchner, 1998; Lavoie, 2001). In spirantization, the same underlying sound is realized as a stop at the beginning of a prosodic constituent, but as a fricative or approximant internal to prosodic constituents. The most common context for the continuant realization is in between vowels, though in some languages it also affects consonants in other positions (Kirchner, 1998; Lavoie, 2001). Figure 1 gives an illustration of spirantization from a speaker of Campidanese Sardinian.

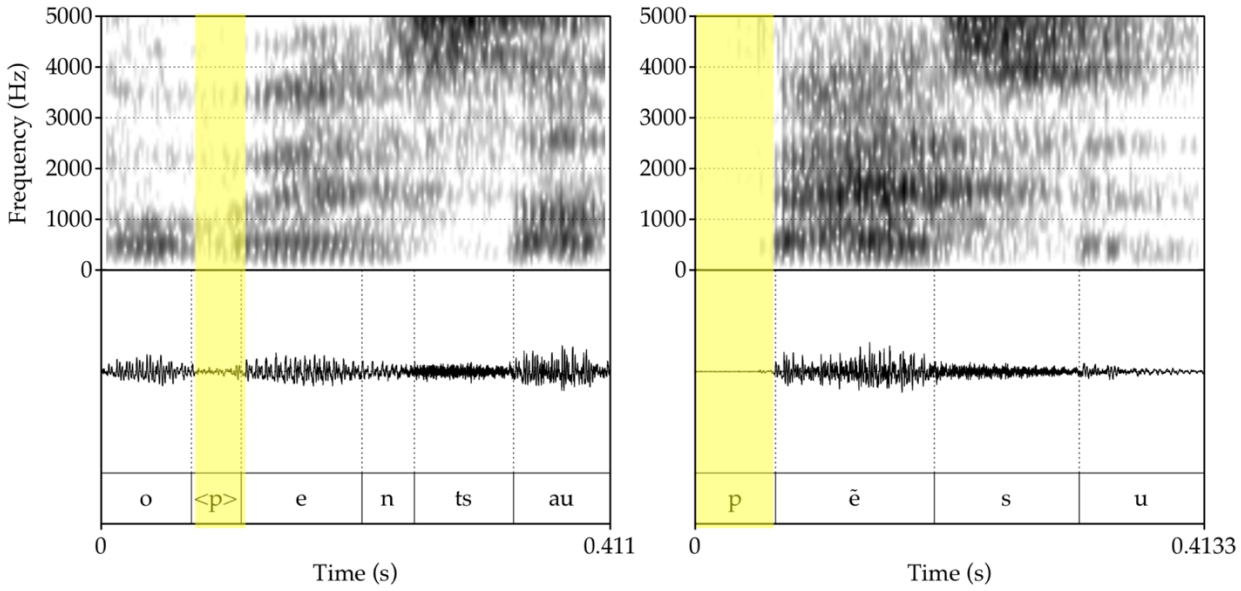


Figure 1. Two tokens of underlying /p/, from forms of the verb /pentsai/ ‘to think’ uttered by a speaker of Campidanese Sardinian. The sounds in question are lightly highlighted. The token on the left follows the final /o/ vowel of the auxiliary verb /ap:o/ and is lenited. The token on the right is utterance initial and is not lenited.

The figure shows two forms of the verb /pentsai/ ‘think’. The phrase-medial form on the left is a past participle that follows the vowel-final auxiliary verb [ap:o] ‘I have’. The initial consonant, lightly highlighted, is realized as a bilabial approximant with formant structure throughout and no burst. The utterance-initial form on the right is the first-person singular present. The initial consonant, also highlighted, is a voiceless stop with a clear burst.

Spirantization is arguably rooted in inherent properties of the human auditory system, and Katz & Fricke (2018) provide evidence that English speakers process spirantization in ways that reflect its prosodic function, despite having little evidence for this function from English itself. In this study, we compare a cross-linguistically common spirantization pattern to a cross-linguistically uncommon (perhaps unattested) one.

Spirantization is not entirely unfamiliar to American English speakers. Non-coronal voiced stops /b/ and /g/ are sometimes realized as approximants in intervocalic position (Lavoie 2001; Warner & Tucker, 2011), though not nearly as reliably as in better-known cases such as Andalusian Spanish (Romero, 1995). American English also differs from other spirantization languages in that the probability of continuant realizations depends more on stress than on the presence of a prosodic boundary (Bouavichith & Davidson, 2014). Nonetheless, Katz & Fricke (2018) show that a typical boundary-conditioned spirantization pattern facilitates word-segmentation for English-speaking adults. They posit that this segmentation effect exists for English speakers, despite its incongruence with their native-language experience, because the segmentation effect is a consequence of general psychoacoustic properties rather than domain-specific linguistic principles. In particular, lenition-fortition patterns tend to place relatively large changes in acoustic parameters at constituent boundaries, and to ensure relative continuity in acoustic parameters internal to constituents (Katz 2016, Keating 2006, Kingston 2008). They can

thus be seen as language-specific instantiations of general Gestalt grouping principles (Wertheimer, 1938).

The purpose of the current study is to test for the spirantization effect in English-speaking school-age children. Specifically, we ask whether children's statistical learning will benefit when the stimuli are more phonetically natural (i.e. comprise a pattern attested in real-world languages, here spirantization) compared to when the stimuli are phonetically unnatural (i.e. comprise a pattern unattested in real-world languages). If the spirantization effect is fundamentally psychoacoustic in nature and independent of linguistic experience, then it necessarily holds for younger and less experienced participants. If we find instead that the effect is absent for children, it would suggest that the original effect in adults is due to some type of acquired knowledge, whether that involves linguistic experience, explicit reasoning and pattern-matching, or some other factor. [1]

Word-segmentation, long-term memory, and experimental design

The second objective of the current study is to examine retention of items from statistical learning paradigms. There are several studies suggesting that items from artificial languages can be retained after the initial exposure period. Artificial languages with conflicting probabilistic patterns can interfere with one another when presented in intermittent blocks (Weiss, Gerfen, &

[1] For the sake of full disclosure: the attempted replication with children was meant to be the first step in a project comparing this population to children with language impairment. Children with language impairment reportedly have difficulty with both word-segmentation (Evans *et al*., 2009) and basic psychoacoustic tasks (Corriveau, Pasquini, & Goswami, 2007; Ziegler, Pech-Georgel, George, & Lorenzi, 2011). Because children showed only marginal learning effects in our study, the study purpose has been adapted accordingly.

Mitchel, 2009) or in immediate succession (Bulgarelli & Weiss, 2016; Gebhart, Aslin, & Newport, 2009). With much longer periods and larger quantities of exposure than those used in laboratory experiments, words acquired through the segmentation paradigm can persist in memory for years (Frank, Tenenbaum, & Gibson, 2013). Adult subjects who are taught novel words in isolation with orthography and semantic referents show lexical competition effects from those words eight to ten months later (Hultén, Laaksonen, Vihla, Laine, & Salmelin, 2010; Tamminen & Gaskell, 2008). In implicit phonological learning, adults still show signs of novel phonotactic constraints one week after training that includes orthography and production tasks (Warker, 2013).

To the best of our knowledge, however, long-term retention in statistical learning tasks has not been tested for school-age children and, given Gomez's (2017) argument from a statistical learning standpoint that there may be different memory systems involved in adults compared to infants, it is not clear that previous retention studies with adults can be generalized to school-age children. Further, the only test using an auditory paradigm like that of the current study (Frank *et al*., 2013) involved roughly 10 hours of exposure over 10 days, whereas a typical laboratory study involves less than one hour on one day. Thus, examining retention within an auditory-only word segmentation paradigm in both school-age children and adults can inform our understanding of the long-term memory processes that are involved in the task and whether children demonstrate adultlike patterns of performance. In this study, we ask whether children and adults retain information from a word segmentation task for at least a month. Consistent with the general Gestalt grouping principles described above, one prediction is that the more phonetically natural artificial language comprising the spirantization pattern will result in better

retention compared to the phonetically unnatural artificial language given that the statistical information may be more easily chunked to facilitate memory (e.g., Christiansen, 2019).

We also hope to provide insight as to whether a within-subject design reasonably can be used in studies of word segmentation. This is important because there are practical benefits to such designs. Recruiting and working with large samples of children is time- and labor-intensive, and low experimental power is an ongoing concern in the infant word-segmentation literature (see Black & Bergmann, 2017; Bergmann *et al*., 2018 for meta-analytical calculations). Within-subject designs greatly increase the power of experiments (e.g. Kirk, 1982; Maxwell & Delaney, 2004). Conceptually, the reasons why are simple. First, one gets twice as much data from each subject, so fewer subjects are needed. And second, any differences between conditions in a within-subject design can be more confidently attributed to the effect of condition than in a between-subjects design, where they may also be due to by-subject variability.

METHOD

Participants

Participants included two groups consisting of 22 children and 36 young adults. The children were recruited from flyers posted around the West Virginia University (WVU) campus, word of mouth, and in-person invitations offered at various after school programs in the local school district. Of the 26 children who were initially enrolled in the study, 24 completed both testing sessions. Two participants' data were excluded from the study due to a computer malfunction during one of the children's sessions and the other child being noncompliant throughout portions of the tasks. Therefore, the final sample included 22 children (14 females) in third through fifth grade whose average age was 9.7 years (range 8.1 – 11.3 years) at the time of enrollment.

Parents reported that all of the children were native English monolinguals and had no history of hearing, vision, or cognitive impairment, nor any other relevant medical concerns. Per parent report, none of the children who were included in the study were currently receiving special education services or speech/language therapy at the time of the study nor in the past. All parents of the participants signed an informed consent, and all children signed an informed assent using procedures approved by the West Virginia University Institutional Review Board. Children received an age-appropriate story book at the end of each study session for their time and effort.

College students were recruited from WVU undergraduate classes and signed an approved informed consent before participating in the study. After completing both study sessions, they received extra credit for participating. Of the initial respondents, 36 (3 males) completed both study sessions. They were 18-26 years of age (M = 19.5, SD = 1.4), monolingual native English speakers, with no self-reported history of hearing, vision, speech, language, or cognitive impairment, nor any other relevant medical concerns. A few exceptions include a participant who reported having ADHD, one participant who reported receiving treatment for dyslexia as a child, and three participants who reported receiving articulation therapy as children. These five adult participants performed within one standard deviation of the mean on the word-segmentation task, and none of them scored in the lowest quartile. Thus, all results below include these participants.

Because the experiment involves cognitive and memory processes, we administered standardized tests of all participants' phonological memory and basic cognitive ability to use as correlates with the word segmentation task. On average, both groups of participants performed within one standard deviation of the normative average (M(SD) = 100(15) for composite scores, M(SD) = 10(3) for subtest scores) on the *Sentence Recall* subtest of the Clinical Evaluation of

Language Fundamentals – Fifth Edition (CELF-5) (Wiig, Semel, & Secord, 2013), the Phonological Memory Composite of the Comprehensive Test of Phonological Processing – Second Edition (CTOPP-2) (Wagner, Torgesen, Rashotte, & Pearson, 2013), and the 2-Subtest IQ of the Weschler Abbreviated Scale of Intelligence, Second Edition (WASI-II) (Wechsler, 2011) (See Table 1). The Phonological Memory Composite of the CTOPP-2 includes the *Nonword Repetition* and the *Memory for Digits* subtests. The 2-Subtest IQ of the WASI-2 includes the *Vocabulary* and *Matrix Reasoning* subtests.

Table 1. Standardized test results.

Test	Children *M* (*SD*)	College Students* *M* (*SD*)
CTOPP-2		
Nonword Repetition	8.0(3.1)	6.9(2.2)
Memory for Digits	9.8(2.4)	10.6(2.5)
Phonological Memory Composite	94.1(11.4)	93.8(11.9)
WASI-II		
Vocabulary	12.0(1.8)	10.7(2.2)
Matrix Reasoning	10.8(2.1)	10.5(2.6)
Full Scale IQ – 2 Composite	108.0(8.6)	103.3(10.5)
CELF-5		
Recalling Sentences	11.1(2.4)	10.1(2.2)

Note: Normative *M*(*SD*) for subtests = 10(3), Normative *M*(*SD*) for composite scores = 100(15).

* - One participant aged 26 years exceeded the normative age range for the CTOPP-2 and CELF-5 subtests, so the data for the max age (CTOPP-2 = 24 years; CELF-5 = 21 years) were used to calculate the participant's scaled and composite scores for these two tests.

Stimuli

The training stimuli used here are identical to those in the spirantization-1 condition of Katz & Fricke's (2018) study on adults. Several aspects of those stimuli, in turn, are based on Frank *et al*., 2010. They were synthesized using the KLSYN-88 speech synthesizer (Klatt & Klatt, 1990), which allows a high degree of control over the fine phonetic characteristics of stimuli. Targets

for the consonants were derived from recordings of a female Madrileño Spanish speaker. The goal was to ensure that our stimuli display the acoustic characteristics of actual sounds involved in spirantization, and Spanish is a canonical example of such a pattern (Romero, 1995). Formant transitions between consonants and vowels were interpolated according to the procedure in the Klatt manual. F0 was held flat at 165 Hz in all syllables. Consonants with their accompanying formant transitions were 120 ms long, with steady-state vowels of 140 ms. Stops had prevoicing for their entire closure duration. Continuants had no noise component; that is, they were realized as sonorants rather than fricatives. Syllables were concatenated together with no intervening silence to make words, and words were concatenated together with no silence intervening to make utterances. Utterances were separated by 800 ms of silence.

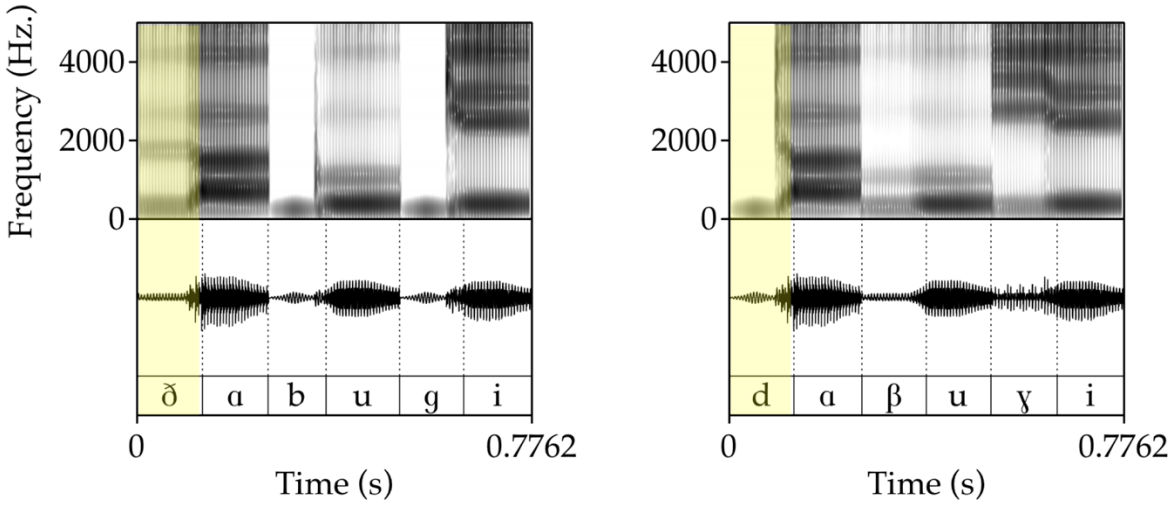


Figure 2. Examples of words used in the experiment, from the anti-spirantization (left) and spirantization (right) conditions. The initial consonants in each word are highlighted for purposes of comparison. The second and third consonants can also be compared to see the difference between stop and approximant realizations.

In the spirantization condition, words featured stop segments initially and approximant consonants medially. The anti-spirantization condition was statistically identical, but the manner of articulation was switched so that initial consonants were approximants and medial ones were stops. This anti-spirantization pattern, to the best of our knowledge, is unattested in real-world languages. Comparing performance on the spirantization and anti-spirantization conditions allows us to isolate the effect of the phonetic content on segmentation, rather than the effect of the mere presence of a phonetic pattern. Example spectrograms and waveforms of one word from each condition are shown in Figure 2.

Note that possible words in each condition are impossible in the other condition. So if subjects retain the pattern they learn first, they should fail to learn the 'opposite' language in the second session and may even favor foils over targets. Both languages had six distinct words, which ranged from two to four CV syllables combining the vowels [a], [i], and [u] with consonants [b], [d], [g], [β], [ð], and [ɣ]. During the exposure period, participants listened to 150 utterances (that is, 150 strings of four words each, concatenated into a continuous acoustic stream) separated by pauses 800 ms in length. The order of words was pseudo-randomized so that the same word was never repeated twice in a row in a single utterance, and the exposure period generally lasted about eight minutes. As an illustration, a possible utterance during the exposure period for the spirantization language, with words visually separated by periods in the transcription, would be [daβuɣi.baði.duɣaβuði.biɣaðu].

Experimental procedure

Following enrollment, the participants completed two separate study sessions that were conducted at least 30 days apart (Children: *M* = 36.5 days between Session 1 and 2, *Range* = 30

– 51 days; Adults: *M* = 35.9 days between Session 1 and 2, *Range* = 30 – 49 days). All study session procedures were administered by trained and supervised student research assistants, and participants completed sessions individually in a classroom at an afterschool program location (i.e., an elementary school) or on campus at West Virginia University.

Session 1 included *Vocabulary* and *Matrix Reasoning* subtests from the WASI-II, *Nonword Repetition* from the CTOPP-2, and a word segmentation task in either the spirantization or the opposite, anti-spirantization, condition. For Session 2, participants completed *Memory for Digits* from the CTOPP-2 and *Recalling Sentences* from the CELF-5 followed by the remaining word-segmentation condition. The word segmentation paradigm was administered on a laptop with headphones using Open Sesame software (Mathôt, Schreij & Theeuwes, 2012) and conditions were counterbalanced by session across subjects. Subtests from the standardized tests were administered according to the standardized procedures described in their respective manuals.

Experimental Task Administration

Participants were seated approximately 16 inches from the center of a laptop and listened to stimuli through Sony Dynamic Stereo MDR-V6 headphones. The presentation volume was set at a comfortable listening level and held constant across subjects. Keyboard responses were recorded using a blue and red sticker over the 'z' and '/' keys of the laptop, respectively.

Participants were told that they would "listen to a made-up language for a few minutes" and then "be asked about the words that appear in that language." Instructions were presented on the screen, but also read aloud to participants. The exposure phase consisted of passive listening for approximately 8 minutes. Instructions to "pay attention to the speaker" because "you will be

388 asked about the words in this language later" remained on the screen during this phase, and
389 participants were instructed to color quietly.
390 After the exposure phase, participants completed a two-alternative forced-choice task in
391 which they were asked which of two strings of syllables (presented acoustically, with no
392 orthographic representation) was "a word in the language [they] just heard." The target in each
393 trial was a word from the language, where most transitions between syllables had a probability of
394 1.0 in the exposure period (three syllables had to be used in more than one word, such that
395 transitions involving these three had probabilities of 0.5 given the surrounding syllables). Foils
396 included sequences that did not occur in the language. In these foils, most syllable sequences had
397 conditional probabilities of 0, though a few sequences had conditional probability 0.2 (sequences
398 across word boundaries in the exposure period). There were a total of six targets and six foils.
399 Each target was heard six times in the testing phase, paired with a different foil each time, for a
400 total of 36 test trials. As an illustration, a subject in the spirantization condition might be asked
401 which one of [biɣaðu] and [βiɣuda] is a word in the language they heard. The former is a word in
402 the language, with transitional probabilities of 1 for the first and second syllables and 0.5 for the
403 second and third. The latter did not occur in the exposure period, and has transitional
404 probabilities of 0 for all syllable sequences.
405 In instructions preceding the forced choice task, participants were told that the questions
406 were hard and that they should do their best to answer correctly. If they didn't know a correct
407 answer, they were asked to guess. As the first item in a pair was played, the instructions to "press
408 the blue key for the first word" appeared on the left-hand side of the screen (in line with the blue
409 sticker on the 'z' key), and when the second test item was played the instructions to "press the
410 red key for the second word" on the right-hand side of the screen (in line with the red sticker on

411 the ‘/’ key). These written prompts remained on the screen with another prompt: “Which word is
412 part of the language you heard?” They were given 4 seconds to respond before a “Timeout”
413 message appeared and the next trial began. The total duration of the experiment including
414 exposure and tests phases was approximately 20 minutes for most participants.
415 The experimental task was identical to that used for adults by Katz & Fricke (2018), with
416 two exceptions. To make the task easier, the foils in our design had transitional probabilities of
417 zero; Katz & Fricke used part-word sequences with non-zero probabilities. And to equalize the
418 frequencies of every item during the test phase, we crossed all 6 targets with all 6 foils; Katz &
419 Fricke paired each target with 5. While the primary motivation for including both children and
420 adult participant groups in the current study was to make direct between-group comparisons in
421 learning, with only these two minor changes in methodology from the Katz & Fricke (2018)
422 study, this also offers a systematic replication of adult participants across the two studies.
423
424 **CHILDREN’S RESULTS**
425 The most notable results here are the lack of differences in learning between different sessions
426 and phonetic conditions, and the weak overall effect of learning. Boxplots of by-subject accuracy
427 are shown in Figure 3, separated by session and condition. We do not include timeout trials in
428 any of the analyses here: they constitute 4.6% of trials (about 2 per subject per session), with 32
429 timeouts in the spirantization condition and 41 in the anti-spirantization condition. In addition, 1
430 trial per subject per session was programmed incorrectly to present 2 foils and 0 targets; these 44
431 total trials are excluded.

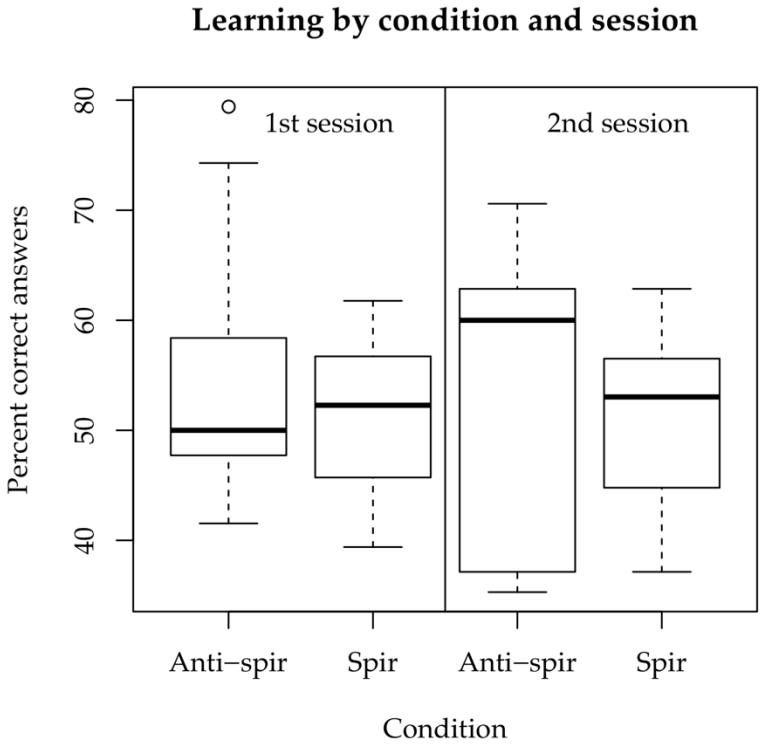


Figure 3. Boxplots of by-subject accuracy in the word-segmentation task by session and phonetic condition for the children.

In 3 of the 4 subsets, median accuracy is in the 50-55% range, and there are no obvious differences by condition or session. The exception is for subjects learning the anti-spirantization language during the second session: their median accuracy is 60%, but this group is also more variable than the others. Pooled across all sessions, we find 26 scores above chance (50%), 17 below chance, and one at chance. These counts differ very little for the first and second sessions.

To investigate the effects of session and condition on learning, we fit logit mixed effects models of accuracy using the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2015). These models estimate the log odds of responding accurately as a function of various random and fixed effects. Our design includes crossed random effects of target word, foil word, and subject. We examined the fixed effects of session, phonetic condition, and centered trial number

(to account for habituation and/or fatigue during the course of the experiment). We attempted to fit a model with the 'maximal' random effects structure first (Barr, Levy, Scheepers, & Tilly. 2013). The presence of by-subject random slopes for centered trial, however, resulted in a random-effects correlation of -1 and a convergence warning, which generally indicate that the model is too complex for the data to which it is fitted. We backed off to a model without those random slopes, which presented no convergence issues. The model summary is shown in Table 2.

Table 2. Summary of logit mixed effects model of accuracy, Experiment 1.

Random effects		**Variance**	***SD***	
Word 1	Intercept	0.17	0.41	
Word 2	Intercept	0.13	0.36	
Subject	Intercept	0.11	0.33	
	Condition: spir	0.22	0.47	
	Session: 2nd	0.03	0.18	
Fixed effects	***β***	**SE**	***z***	***p***
Intercept	0.26	0.20	1.30	0.19
Condition: spir	-0.15	0.27	-0.56	0.58
Session: 2nd	-0.04	0.14	-0.31	0.76
Trial (centered)	0.004	0.005	0.69	0.49

Spir = spirantization.

As suggested by the plot in Figure 3, neither session nor phonetic condition has a robust effect on learning. The effect of trial is also negligible. The intercept here represents accuracy above chance in the anti-spirantization condition on the first session: while it is the largest fixed effect in this model, the effect of learning here is small and not particularly robust. In other words, this study does not find strong evidence that children can do the word-segmentation task.

While this study estimates learning using a mixed-effects regression model that accounts for random variables in a principled way, many studies in this area instead use one-sample t-tests

over by-subject summary statistics (e.g. Evans *et al*., 2009; Saffran *et al*., 1997). Because the results here are different than those earlier studies, it is worth asking whether those differences stem from statistical procedures. In our within-subject design, there is no perfectly equivalent way to perform a one-sample *t* test. We pooled accuracy across both conditions. Median accuracy is 51% and mean is 53%. The standard deviation is about 7%. The results from a one-tailed *t* test indicate that these data are inconsistent with a population of by-subject means normally distributed around chance (one-tailed $t(21) = 2.03$, $p = 0.03$). For the sake of comparison, a logit mixed-effects model with random effects but no fixed effects of condition, session, or trial returns a learning effect of 1.24 standard errors, $p = 0.22$. This suggests that modeling assumptions do indeed make a substantial difference for the interpretation of our data.

The upshot of this discussion is that subjects on average performed just slightly better (1-2 more correct answers in a 36-item test) than chance. The significance of this effect depends on the assumptions embedded in the statistical model: models that attempt to generalize across items return smaller and less robust estimates. *P*-values for the general effect of learning range from 2-22% depending on the statistical model used.

As a follow-up, we explored by-subject variability in these data in an attempt to account for the unexpectedly weak performance of this sample. One possibility is that children with high language aptitude or IQ perform well on the task, but other children have trouble with it. Table 3 shows Pearson correlations between by-subject pooled accuracy across the two word-segmentation conditions and standardized test results (plus age).

None of these attributes account for more than 10% of the variance in the word-segmentation task, and the correlations here are non-significant. This provides no support for the

idea that children with some well-defined set of characteristics are able to perform the word-segmentation task to the exclusion of others.

Table 3. Pearson correlations between word-segmentation accuracy pooled across both sessions, and various measures of aptitude and age for children.

Item	Pearson *r*	*p*
Age	0.15	0.52
NWR	0.08	0.72
Sentence recall	0.19	0.39
Digit recall	-0.31	0.16
IQ	0.15	0.50

NWR = Nonword repetition.

A weaker hypothesis is that some children have a stable and consistent ability to perform the word-segmentation task, but this ability is not being captured by the standardized tests used here. To check this, we calculated the split-half reliability of the word-segmentation task across conditions, splitting the data into odd and even trials. The split-half reliability is $r = 0.27$, $t(20) = 1.24$, $p = 0.23$. This suggests that the word-segmentation task is barely measuring any stable property of children at all: only about 7% of the variance in performance can be attributed to subject identity.

DISCUSSION – CHILD OUTCOMES

This study failed to extend to children the finding that spirantization aids word segmentation for adults (Katz & Fricke, 2018). This is unsurprising, because the study only found a marginal effect of learning in any phonetic condition or experimental session. When results are pooled across conditions and sessions, a slightly more robust effect of learning emerges, but the magnitude and certainty associated with that effect vary with different statistical assumptions.

Furthermore, the task shows little sign of external or internal validity, as assessed by correlation with test scores and split-half reliability, respectively.

The learning effect here is smaller and less robust than most previous results in this literature. This is addressed in detail in the general discussion section. Before we start to question the general robustness of word-segmentation ability in this age group, however, it is important to ask whether these weak results might be a consequence of the particular materials and methods we used here. Many previous studies in this age group have used the same stimuli and design as Saffran *et al*. (1997), consisting of repeating sequences of six 3-syllable words (e.g. Evans *et al*., 2009; Mainela-Arnold & Evans, 2014; Mayo & Eigsti, 2012). Our experiment differed from this setup in several ways. First, there is the within-subject design itself, although the results suggest this made little difference in terms of performance. Our materials are also different: we included systematic phonetic patterns in our 'words', varied their syllable count, and separated them into 4-word 'utterances' during the exposure phase. Our stimuli are modeled after Spanish sounds, which may make the task more difficult for English speakers (Perruchet & Poulin-Charronnat, 2012; Finn, Hudson Kam, Ettlinger, Vytlacil, & D'Esposito, 2013). Finally, the phonetically-detailed and tightly controlled synthesis procedure used here is quite different from the Saffran *et al*. (1997) stimuli, which come from a commercial text-to-speech product. For any of these reasons, our task may have been more difficult than in previous studies, many of which share a single set of stimuli and procedures.

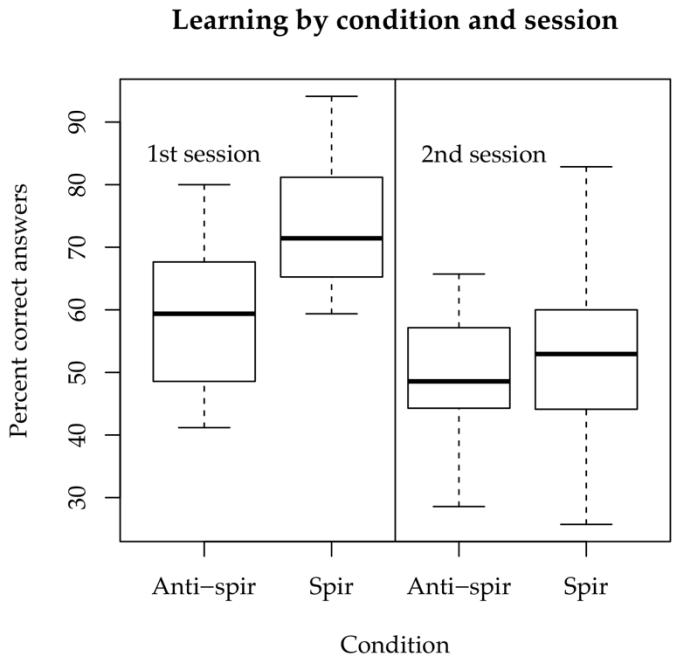
For this reason, we included adults as a second participant group in the study. There is no doubt that adults are capable of understanding and performing the word-segmentation task, and learning effects in this age group tend to be quite large and unambiguous. In the study used as a model for this one, for instance, college students attained 64% accuracy in one phonetic

condition and upwards of 80% in the other, with materials very similar to those used here. Our reasoning for the adult participant group is as follows: if college students don't show significant learning with the procedure, the most likely hypothesis is that the word-segmentation task was overly difficult. If college students perform similarly to past studies, on the other hand, we can conclude that weak performance observed in children was due to the subject group, not the properties of the experiment.

ADULT RESULTS

The college students in this study were much more successful on the word segmentation task than the children. Timeouts, not included in the results shown here, constituted just under 1% of trials. The same coding error from the child data resulted in 1 erroneous trial per subject per session, which was discarded. Boxplots of by-subject accuracy, separated by session and condition, are shown in Figure 4.

For the first session, performance is higher on the spirantization condition than the anti-spirantization condition, as reported by Katz & Fricke (2018). Even for the anti-spirantization condition, about two thirds of the distribution is above chance (50%). Mean accuracy in the first session is about 74% (*SD* = 13) in the spirantization condition and 59% (*SD* = 11) in the anti-spirantization condition. Performance is uniformly worse in both conditions on the second session, with substantial portions of the distribution at or below chance.



548

549 Figure 4. Boxplots of by-subject accuracy in the word-segmentation task by session and
550 condition for the adults.

551

552 A summary of our final regression model is shown in Table 4. Random effects turned out to be
553 complicated for this model. The model with maximal random-effects structure did not converge.
554 By-subject slopes for both condition and session contributed substantially to model fit, but
555 incorporating both into the same model resulted in convergence failure and degenerate Hessian
556 warnings, even after rescaling the trial variable and changing optimizers. Because random slopes
557 for condition resulted in a lower Bayesian Information Criterion than those for session, we chose
558 to keep condition in the final model. Note that even though this process was complicated and
559 involved a lot of models, all of these models (even the ones that failed to converge) produced
560 sensible fixed effects estimates, none of which differed qualitatively from the final model below.
561 These basic fixed effects patterns thus appear to be robust to a variety of different assumptions
562 about random effects structure.

Table 4. Summary of logit mixed effects model of accuracy, Experiment 2.

Random effects		**Variance**	***SD***	
Word 1	Intercept	0.13	0.37	
Word 2	Intercept	0.08	0.27	
Subject	Intercept	0.10	0.32	
	Condition: spir	0.45	0.67	
Fixed effects	***β***	**SE**	***z***	***p***
Intercept	0.42	0.18	2.35	0.02
Condition: spir	0.73	0.28	2.66	< 0.01
Session: 2nd	-0.42	0.16	-2.63	< 0.01
Trial (centered)	-0.01	0.004	-2.16	0.03
Interaction: Spir * 2nd Session	-0.50	0.28	-1.77	0.08

Spir = spirantization.

The first result of interest is that performance in the spirantization condition is substantially better than the anti-spirantization condition, especially for the first session. This broadly replicates the findings of Katz & Fricke (2018). A second major pattern is that performance is lower in the second session. This effect is larger for the spirantization condition, although the interaction term doesn't quite reach the alpha level of 5%.

For the children in this study, performance on the word-segmentation task did not consistently track standardized test scores nor subject identity. We examined these effects for the adults as well. Because of the substantial carryover effect between sessions in these data, we used accuracy in the first session only to approximate 'uncontaminated' word-segmentation performance. Correlation with standardized tests (and age) are shown in Table 5.

As with the children in this study, there is little correlation between age or test scores and performance in the experiment. A possible exception is the nonword repetition test, which correlates at least moderately with word-segmentation performance. That said, *p*-values are

unreliable when conducting such a large number of tests and it is entirely possible that a correlation of this magnitude could arise by chance.

Table 5. Pearson correlations between word-segmentation accuracy in the first session, and various measures of aptitude and age for adults.

Item	**Pearson *r***	***p***
Age	0.04	0.81
NWR	0.36	0.03
Sentence recall	-0.05	0.77
Digit recall	0.08	0.65
IQ	-0.02	0.89

NWR = Nonword repetition.

Unlike the children, college students were reasonably internally consistent in their performance. Split-half reliability in the first session for adults is $r = 0.61$, $t(34) = 4.44$, $p < 0.001$. While this would be considered mediocre for a psychometric or diagnostic test, it still indicates that the task is capturing a substantial amount of stable information about individuals.

We found a fairly large decrease in accuracy in the second session relative to the first. Inspection of Figure 4 shows that this includes a decrease at the bottom of the distribution: while only 1 in 6 adult participants performed at or below chance in the first session, nearly half did so in the second session. This includes a number of subjects with accuracy unusually far below chance: the minimum goes from 41% correct in the first session to 26% in the second. In that second session, 7 subjects scored 40% or lower, with 2 below 30%. For perspective, the binomial test suggests that random guessing in a 36-trial Bernoulli process should produce accuracy below 30% for less than 3% of subjects. We think these data are consistent with the idea that a subset of subjects actively prefer the foils over the targets during the second session.

To address the alternative hypothesis that a decrease in effort in the second session resulted in chance performance for some subjects but no active preference for foils, a subset of

adult participants (n = 26) were asked to rate their level of effort on the word segmentation task after each of the experimental tasks were completed. On a scale of 1 to 5 (1 = "I did not try at all," 5 = "I tried really hard to answer the questions well"), students were asked to "rate [their] level of effort on the questions about the made-up language at the end of the session." The difference in ratings from Session 1 to 2 is in the opposite direction of the proposed decrease in effort ($M(SD)_1$ = 4.6(0.5), $M(SD)_2$ = 4.7(0.5); two-tailed paired-samples t test: $t(25) = -1.81$, p = 0.08). For this reason, we do not think it is likely that the accuracy effect is due solely to increased guessing or decreased effort during the second session.

DISCUSSION – ADULT OUTCOMES

The college students performed much closer to expectations on the word-segmentation task, at least in the first session. There were substantial and robust learning effects for both conditions in the first session, with accuracy in the spirantization language significantly higher than the anti-spirantization language. The spirantization advantage is smaller in the second session, where performance declined precipitously for both conditions.

This broadly replicates Katz & Fricke's (2018) results, obtained with similar stimuli and methods, also testing college students. And it suggests that the problem with the children's results was not the stimuli or procedure: it is possible to get robust learning results with these materials. That said, overall accuracy here, even in the first session, is somewhat lower than reported by Katz & Fricke: subjects in that study averaged 82% accuracy in the spirantization condition and 64% in the anti-spirantization condition (the numbers here are 74% and 59%, respectively). It is not possible to explain this difference with only evidence from the two studies in question. It may be normal sampling variation, may pertain to differences in the test procedure

(though these are very minimal), or may be due to group differences: subjects were sampled from a UC Berkeley language-acquisition class in the previous study, from WVU Introduction to Communication Sciences & Disorders and Public Speaking classes in this one.

The large decline in performance during the second session is consistent with the hypothesis that some subjects retained information about the words from the first session during the second session, more than a month later. This would hold regardless of which condition was completed first, because a target word in one condition is necessarily illegal in the other condition. If participants answer questions in the second session based on information retained from the first session, they should perform below chance. While it's hard to determine conclusively whether specific participants' below-chance results are due to this carryover effect or just random guessing, it is worth noting the most extreme low scores (below 40% accuracy) all occurred in the second session. That said, we cannot rule out the possibility that subjects were simply less focused or cooperative during the second session, and that the low extrema result from random guessing. Adult participants' self-reported effort levels, however, suggest that if anything they tried *harder* during the second session.

GENERAL DISCUSSION

Language acquisition research has utilized statistical learning paradigms to demonstrate implicit learning of both phonological and morphosyntactic structure of language. Within minutes, the passive listener is able to utilize regularities in an artificial language to infer features such as word boundaries or grammaticality. Previous research has shown that language-specific acoustic patterns interact with such learning in interesting ways. The focus in this study was the role of spirantization, a cross-linguistically common phonetic pattern, in facilitating learning in children.

An ancillary question, largely independent from the first and motivated by methodological concerns, was whether items from the experiment were retained for at least a month. This is a greater retention time frame than what has been previously tested within a short, laboratory implementation of the word-segmentation paradigm.

To address the goals of the study, we exposed third through fifth graders to two artificial languages, in sessions at least one month apart. One artificial language comprised the spirantization pattern (stops word-initially, approximants word-medially) and the second comprised an anti-spirantization pattern (approximants word-initially, stops medially). The results showed no benefit from spirantization nor long-term retention, though this lack of finding must be considered in the context of the children's overall learning during the task: group performance across conditions on the two-alternative forced choice task was barely above chance, thus demonstrating little or no learning at all regardless of phonetic pattern.

To rule out task design as an explanation for minimal learning in the children, the same word-segmentation paradigm was administered to a group of college students, since previous work has shown that spirantization facilitates statistical learning in this age group (Katz & Fricke, 2018). The findings from this study replicate the previous work: college students demonstrated learning in both conditions, with improved performance in the spirantization condition compared to the anti-spirantization one. This suggests that the domain-general acoustic properties of languages play at least some role in helping to identify constituents within a language, even if those acoustic properties mismatch to some extent with one's native language.

669 *Children's low performance*

670 Although the *existence* of word-segmentation abilities based on statistical learning are robustly
671 replicated and not in doubt, there has been some concern over the size of such effects, their
672 generality across individuals, and the reliability of the word-segmentation task as a measure of
673 individual abilities. This is true for both infants and for the school-aged children we studied here.

674 In the infant literature, Black & Bergmann's (2017) meta-analysis concludes that there
675 are probably 'real' average effects at the population level, but also shows tentative evidence for
676 an effect of publication bias in one narrow part of the literature (their figure 1), and a general
677 picture of the literature (their figure 3) where studies with many subjects don't seem to be any
678 more consistent in their findings than studies with very few subjects. Bergmann *et al*. (2018), in
679 a more extensive meta-analysis of language acquisition literature in general, find a significant
680 negative correlation between effect sizes and sample sizes in infant word-segmentation, which
681 may indicate researcher degrees of freedom in study design. It is difficult to interpret the infant
682 literature as a whole because preferences switch back and forth from novel stimuli to familiar
683 ones in different experiments, and sometimes within the same experiment (e.g. Graf-Estes &
684 Lew-Williams 2015).

685 There are far fewer studies of word-segmentation in school-aged children, but recent
686 developments suggest some cause for concern. Using a 2-alternative forced choice task, Raviv &
687 Arnon (2018) report a smaller learning effect (about 55% correct on average) and more
688 variability than the initial, smaller studies in this age group. A follow-up study by Arnon (2019)
689 finds that word-segmentation performance is an inconsistent and unreliable measure of
690 individual variability in this age group. For adults, the consistency and reliability of the word-
691 segmentation task is higher, but still falls short of standards for psychometric tests (Siegelman,

Bogaerts, & Frost, 2017). Lammertink, Van Witteloostuijn, Boersma, Wijnen, & Rispens (2019) report that children can't do a 2-alternative forced choice task for a somewhat different type of artificial grammar-learning experiment. The current study reported here can thus be seen as independent, converging evidence that effects in this age-group are not as general nor robust as was initially believed.

These concerns are compounded by our finding that the interpretation of the data depends on whether or not by-item variance is modelled. Most of the earlier studies in this age-group, which tended to produce larger and/or more robust learning effects, were based on the same set of 6 words and 36 2-alternative forced choice trials as the original studies by Saffran, Aslin, & Newport (1998). These studies assessed learning using a *by-subject* one-sample t-test (or equivalent ANOVA for more complex designs), which licenses inferences across people using the particular stimuli in question, but does not license inferences across stimulus properties in general (see e.g., Max & Onghena, 1999 for an extended explanation of this 'language as fixed effect fallacy'). Raviv & Arnon (2018), Arnon (2019), and the current study all used different stimuli, all included by-item variance in statistical models, and all found smaller learning effects. In the current study, we found that the average learning effect went from 'non-significant' to 'significant' when a mixed model incorporating random effects of item and subject was switched out for a within-subject paired-sample t-test. This suggests that idiosyncratic stimulus properties are potentially a major factor in the size and robustness of learning effects, and that explicitly modelling such variables is crucial to drawing reliable conclusions. Siegelman *et al.* (2018) offer converging evidence from adults that the properties of individual stimulus items exert an enormous effect on word-segmentation results, and that the task differs from non-linguistic statistical learning in this regard.

Between-group differences and development

One interpretation of the different outcomes between children and adults is that the requirements for learning change across development. It is possible that the structural elements of our word-segmentation paradigm were sufficient to facilitate learning in adults, but not sufficient for children. In Plante and Gomez's (2018) review of the clinical relevance of statistical learning, they report several practical implications from the statistical learning literature to facilitate language learning. For example, they note the regularity principle in which frequently occurring target forms, as well as targets that are presented consistently across sentences, can facilitate word learning. Similarly, the variability principle states that high variability for nontarget items promotes learning. Yet another point that the authors note is that correct productions of the targets can facilitate learning, perhaps because the retrieval process required for production helps to encode the target in memory.

While Plante and Gomez's review is framed in the context of how to apply principles from statistical learning to the treatment of children and adults with developmental language disorders, we postulate here that these different implicit and explicit factors may be more or less critical in different learning contexts and at different stages of development. Based on the premise that statistical learning taps into basic memory and learning systems (Christiansen, 2019), context- and age-related differences in any memory and learning process could conceivably be reflected in statistical learning performance as well. One example is modality-based learning differences in children 5-12 years old in which age affected learning in the visual domain but not in the auditory domain. Of course, these findings need to be considered among other potential stimulus-based factors that have also resulted in different performance outcomes (e.g. linguistic versus non-linguistic stimuli; Raviv & Arnon, 2018; Arnon, 2019). Given that the

738 limited number of statistical learning studies that include pre-adolescent, school-aged children
739 report mixed results in children's ability to demonstrate learning, there is a need for future work
740 to systematically vary the components of training (i.e. familiarization) to better understand what
741 maximally benefits learning at different age points.

742 Another important consideration in interpreting the different outcomes between children
743 and adults is the stability of children's performance. As was apparent from the box plot in Figure
744 3, the children's individual performance on our task was highly variable, and the lack of internal
745 consistency from the split-half reliability analysis further confirms that the word-segmentation
746 task here is not assessing a stable property in children. These findings persisted even when the
747 children's individual performance on standardized cognitive-linguistic measures was factored
748 into the analysis. These results are not surprising given recent work that has reported low
749 psychometric validity in statistical learning (e.g., Siegelman et al., 2017; Arnon, 2019), as well
750 as the general fact that children often display higher response variability than adults in speech-
751 related tasks. The psychometric properties are of particular concern as this field attempts to take
752 years of outcomes from group studies and apply it to the examination of individual differences in
753 statistical learning and language ability. As noted by Siegelman et al. (2017), "if a task does not
754 reliably tap the theoretical construct it is supposed to tap (in our case, a postulated individual
755 capacity in [statistical learning]), its explanatory adequacy remains empty" (p. 419).

756 Both Arnon (2019) and Siegelman et al. (2017) suggest that poor psychometric stability
757 in statistical learning tasks could be due to the way that learning is measured. With regard to
758 children specifically, it has been suggested that the 2-alternative forced choice task might be too
759 difficult for them because it requires explicit decision-making and metalinguistic skills that are
760 not fully developed in children (Lammertink *et al.,* 2019). Other measures, such as more implicit

online reaction time measures (e.g. Lammertink et al., 2019; Misyak, Christiansen, & Tomblin, 2010) or the statistically induced chunking recall (SICR) task (Christiansen, 2019) show promise to improve the reliability of the statistical learning paradigm, and may offer an improved way of making age-related comparisons.

While improving measurement is a possibility for increasing the validity of statistical learning tasks, another way to improve the detection of learning effects would be to increase sample sizes. For instance, if the true underlying effect size for children is close to the 0.45 standard deviations reported here, then 31 subjects would be required to have an 80% probability of detecting a 'significant' effect using a one-sided, one-sample *t*-test over by-subject means. This is a larger sample than that of the current study, and larger than most of the other studies we've found in this age group. That said, such statistical tests ignore variability between items, don't license inferences to stimuli beyond those used in a particular experiment, and do not address concerns about internal and external validity. To determine which stable properties of individuals affect the outcome of statistical learning experiments, increasing the sample size will not be sufficient. Increasing the number of trials *per subject* could help, but this is unlikely in practice if 30-50% of children this age simply can't do the experimental task.

In sum, school-age children perform differently than adults on statistical learning tasks, demonstrating learning to a lesser degree (or not at all) and unstable patterns of performance. Although there is a history of work that has shown learning at the group level, the lack of psychometric stability makes it difficult to examine individual differences or more fine-grained adaptations to statistical learning paradigms (e.g. effects of phonetic patterns such as the spirantization pattern examined in this study). Future work could more systematically assess both

783 the training and testing phases of statistical learning to better understand which factors are most
784 important for learning in this age group and to improve the stability of children's performance.
785
786 *Long-term retention*
787 Our study also examined long-term memory effects within the word-segmentation paradigm.
788 Previous work shows some evidence that adults are able to retain, over relatively long timespans,
789 specific wordforms and sound patterns learned in production experiments. For word-
790 segmentation, extended training over 10 days results in retention for years, but this is much more
791 exposure than a one-off laboratory experiment. The current study, in which adult participants
792 completed two sessions at least thirty days apart, strongly suggests that phonological learning
793 affected the second session results more than a month later. This change in performance in the
794 second session was not a factor of perceived effort, suggesting that there may be learning
795 competition effects as a result of long-term memory retention from the Session 1 condition. This
796 is notable in part because our study involved no orthographic representations, no meaning or
797 referents associated with novel items, and no production of the novel items. Previous studies had
798 only shown retention of phonologically learned patterns for up to a week, even with all of the
799 activities above.
800 The findings also showed a larger retention effect for the spirantization condition
801 compared to the anti-spirantization condition. Although this should be interpreted cautiously
802 because performance in the anti-spirantization condition was already lower in Session 1, if this
803 effect were true it could indicate that the acoustic parameters tested here facilitate memory-based
804 processes, such as chunking. Practically speaking, the learning competition effects observed

provide cautionary evidence against using a within-subject design to evaluate competing conditions of a word-segmentation paradigm.

Limitations

There are several factors regarding the participant groups that should be taken into consideration. While the adult participants and the children's parents were asked to report whether they or their child, respectively, had any cognitive deficits, non-language learning difficulties, or relevant medical concerns, they were not directly asked about ADHD nor were they screened for it. Similarly, participants (or their parents) were asked to report on hearing history, but they were not directly screened for hearing. Both groups on average performed within one standard deviation of the normative average on standardized tests that require both adequate hearing and sustained attention, suggesting that it is unlikely that these abilities were significant confounding factors in the study. Nevertheless, given the multiple trials during the testing phase of the experiment which necessitated sustained attention for auditory stimuli, future work should take both attention and hearing ability into consideration.

Another point of consideration is the environment and time of day for the study. Approximately two-thirds of the children completed the study in an elementary school during afterschool care (separated from the other afterschool children) and the other children and adult participants completed the study on the university campus. Although all participants were wearing headphones during the experimental tasks, the different locations are each susceptible to various idiosyncratic ambient noises that could have confounded performance, particularly during the standardized testing when headphones were not worn. To further assess whether ambient noise was a significant confounding issue, we compared the mean standard scores of the

children who completed the study at their afterschool care sites (n = 16) to the scores of the children who completed the study on the university campus (n = 6) using independent sample t-tests. If the listening condition in one environment was negatively impacting performance compared to the other, we would expect to see a consistent pattern of decreased performance across the battery of standardized tests for the children in that environment. Of the five component tasks (the two other scores are composites), two show differences of less than 1/10 of a normative SD between sites. The remaining three show differences of 0.25-1 SDs, but in different directions: two show better scores at the university site, one shows better scores at the school sites. There were no statistically significant differences between the two groups on six of the seven scaled and composite scores (range of p = 0.12 - 1.0); one score was significantly different (p = 0.04), though this is unsurprising given the multiple comparisons examined here. Thus, there does not seem to be a consistent difference in the pattern of performance that could be easily attributed to listening condition. Another consideration is the time of day at which the study sessions were completed. The children completed their study sessions after school when fatigue could have affected performance, whereas the adult participants had more flexibility in scheduling the sessions throughout the day. While there are no obvious differences in average scaled scores on the standardized testing to suggest that the children as a group were significantly negatively impacted by the environment or time of study compared to the adult participant group, these factors could be controlled in future work to rule them out as possible confounds.

CONCLUSION

School-age children's word learning, as measured by a two-alternative forced choice task during a word-segmentation paradigm, was too unstable to identify possible fine-grained phonetic factors that facilitate word learning. Future work systematically could explore a number of

852 factors pertaining to the training and testing phases of learning in order to improve both the
853 children's ability to learn as well as the stability of the task. In contrast, adults demonstrated
854 learning which benefited from modifying a nonnative psychoacoustic feature of the artificial
855 language and was retained for at least a month. Conducting studies of this nature is critical for
856 understanding implicit learning in children and adults, how this learning changes over time, and
857 how to provide maximally beneficial language learning opportunities.
858
859 **CONFLICTS OF INTEREST**
860 The authors have no conflicts of interest.
861
862 **FUNDING**
863 This project was funded in part by Research and Scholarship Advancement Award R883 from
864 West Virginia University to Jonah Katz and Michelle Moore.
865
866 **ACKNOWLEDGMENTS**
867 Many thanks to the research team in the Language and Literacy Lab; to Julia Hamilton, Director
868 of Extended Day Activities; to the site coordinators and volunteers involved with Mountaineer
869 Boys and Girls Club and Afternoon Adventures; to Alex Cristia for helpful discussion; and to the
870 participants and their families for all of their contributions of time and effort to this project.
871
872 **REFERENCES**
873 Arnon, I. (2019). Do current statistical learning tasks capture stable individual differences in
874 children? An investigation of task reliability across modality. *Behavior Research*

Methods. DOI: 10.3758/s13428-019-01205-5

Bagou, O., Fougeron, C., & Frauenfelder, U. (2002). Contribution of prosody to the segmentation and storage of "words" in the acquisition of a new mini-language. Presented at Speech Prosody, Aix-En-Provence.

Barr, D., Levy, R., Scheepers, C. & Tilly, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255-278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015a). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Bergmann, C., Tsuji, S., Piccinini, P., Lewis, M., Braginsky, M., Frank, M. & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development, 89*(6), 1996-2009.

Black, A., & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th annual conference of the Cognitive Science Society* (pp. 124–129). Austin, TX: Cognitive Science Society.

Bouavichith, D. & Davidson, L. (2013). Segmental and prosodic effects on intervocalic voiced stop reduction in connected speech. *Phonetica, 70*, 182-206.

Browman, C. & Goldstein, L. (1990). Tiers in Articulatory Phonology, with some Implications for Casual Speech. In Kingston & Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (341-397). Cambridge, UK: Cambridge Univ. Press.

Bulgarelli, F., & D. Weiss. (2016). Anchors aweigh: The impact of overlearning on entrenchment effects in statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42(10),* 1621-31.

Chomsky, N. (1955). *The logical structure of linguistic theory*. MIT Humanities Library. Microfilm. Published in 1977 by Plenum.

Cohen Priva, U. (2017). Informativity and the actuation of lenition. *Language,* 93(3), 569-597.

Cohen Priva, U, & Gleason, E. (2020). The causal structure of lenition: A case for the causal precedence of durational shortening. *Language,* 96(2), 413-448.

Corriveau, K., Pasquini, E., & Goswami, U. (2007). Basic auditory processing skills and Specific Language Impairment: A new look at an old hypothesis. *Journal of Speech, Language, and Hearing Research, 50*, 647–666.

Christiansen, M. H. (2019). Implicit Statistical Learning: A Tale of Two Literatures. *Topics in Cognitive Science, 11(3),* 468–481. https://doi.org/10.1111/tops.12332

Diehl, R., Lotto, A., & Holt, L. (2004). Speech perception. *Annual Review of Psychology, 55*, 149–179.

Evans, J., Saffran, J., & Robe-Torres, K. (2009). Statistical learning in children with Specific Language Impairment. *Journal of Speech, Language, and Hearing Research, 52*, 321–335.

Finn, A., Hudson Kam, C., Ettlinger, M., Vytlacil, J., & D'Esposito, M. (2013). Learning language with the wrong neural scaffolding: The cost of neural commitment to sounds. *Frontiers in Systems Neuroscience*, 7, Article 85.

Frank, M., Goldwater, S., Griffiths, T., & Tenenbaum, J. (2010). Modeling human performance in statistical word segmentation. *Cognition, 117*, 107-125.

Frank, M. C., Tenenbaum, J. B., & Gibson, E. (2013). Learning and long-term retention of large-scale artificial languages. PloS one, 8(1), e52500.

Frost, R., Monaghan, P., & Tatsumi, T. (2017). Domain-general mechanisms for speech segmentation: The role of duration information in language learning. *Journal of Experimental Psychology: Human Perception and Performance, 43*(3), 466-476.

Frost, R., Armstrong, B., & Christiansen, M. (2019). Statistical Learning Research: A critical review and possible new directions . *Psychological Bulletin , 145*(12), 1128-1153.

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

Goldstone, R.L. (1998). Perceptual learning. *Annual Review of Psychology,* 49, 585-612.

Gordon, M. & Munro, P. (2007). A phonetic study of final vowel lengthening in Chickasaw. *International Journal of American Linguistics,* 73, 293-330.

Graf Estes, K., & Lew-Williams, C. (2015). Listening through voices: Infant statistical word segmentation across multiple speakers. *Developmental Psychology, 51*, 1517-1528.

Harris, J. (2003). Grammar-internal and grammar-external assimilation. In M.J. Solé, D. Recasens, & J. Romero (eds.), *Proceedings of the 15th International Congress of Phonetic Sciences*. Barcelona: Futurgraphic. 281-284.

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, 31, 190-222.

Hultén, A., Laaksonen, H., Vihla, M., Laine, M., & Salmelin, R. (2010). Modulation of brain activity after learning predicts long-term memory for words. *The Journal of Neuroscience, 30*(45), 15160-15164.

Johnson, E. & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory & Language, 44*(4), 548-567.

Johnson, E. K., & Seidl, A. H. (2009). At 11 months, prosody still outranks statistics. *Developmental Science*, *12*(1), 131–141.

Karuza, E. A., Newport, E. L., Aslin, R. N., Starling, S. J., Tivarus, M. E., & Bavelier, D. (2013). The neural correlates of statistical learning in a word segmentation task: An fMRI study. *Brain and Language*, *127*(1), 46–54.

Katz, J. (2016). Lenition, perception, and neutralisation. *Phonology,* 33(1). 43-85.

Katz, J. & Fricke, M. (2018). Auditory disruption improves word segmentation: a functional basis for lenition phenomena. *Glossa, 3*(1), 38.

Katz, J., & Pitzanti, G. (2019). The phonetics and phonology of lenition: A Campidanese Sardinian case study. *Laboratory Phonology, 10*(1), 1-40.

Kim, S. (2004). *The Role of Prosodic Phrasing in Korean Word Segmentation*. PhD thesis, University of California, Los Angeles.

Kirchner, R. (1998). *An effort-based approach to consonant lenition*. PhD thesis, University of California, Los Angeles.

Kirk, R.E. (1982). *Experimental design: Procedures for the behavioral sciences.* Belmont, CA: Wadsworth.

Klatt, D. & Klatt, L. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87*, 820–857.

Lammertink, I., Van Witteloostuijn, M., Boersma, P., Wijnen, F., & Rispens, J. (2019). Auditory statistical learning in children: Novel insights from an online measure. *Applied Psycholinguistics, 40(2)*, 279–302. https://doi.org/10.1017/S0142716418000577

Lavoie, L. (2001). *Consonant Strength: Phonological Patterns and Phonetic Manifestations*.

New York: Garland.

Mainela-Arnold, E., & Evans, J.L. (2014). Do statistical segmentation abilities predict lexical-phonological and lexical-semantic abilities in children with and without SLI? *Journal of Child Language, 41,* 327-351.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314-324.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah, N.J: Lawrence Erlbaum Associates.

Mayo, J. & Eigsti, I. (2012). Brief report: A comparison of statistical learning in school-aged children with High Functioning Autism and typically developing peers. *Journal of Autism and Developmental Disorders, 42*, 2476-2485.

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010). On-Line Individual Differences in Statistical Learning Predict Language Processing. *Frontiers in Psychology, 1*. https://doi.org/10.3389/fpsyg.2010.00031

Nespor, M. & Vogel, I. (1986). *Prosodic Phonology*. Dordrecht: Foris.

Onnis, L., & Thiessen, E. (2013). Language experience changes subsequent learning. Cognition, 126(2), 268-284.

Perruchet, P & Poulin-Charronnat, B. (2012). Beyond transitional probability computations: Extracting word-like units when only statistical information is available. *Journal of Memory and Language, 66,* 807-818.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Mass.

Plante, E., & Gómez, R.L. (2018). Learning without trying: The clinical relevance of statistical

learning. *Language, Speech, and Hearing Services in Schools, 49*, 710-722.

Plante, E., Gómez, R. & Gerken, L. (2002). Sensitivity to word order cues by normal and language/learning disabled adults. *Journal of Communication Disorders, 35,* 453– 462.

Polka, L. & Sundara, M. (2003). Word segmentation in monolingual and bilingual infant learners of English and French. In M.J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of ICPhS 15,* pp. 1021-24. Barcelona: Caudal.

Potter, C., Wang, T., & Saffran, J. (2017). Second language experience facilitates statistical learning of novel linguistic materials. *Cognitive Science, 41,* 913-927.

Raviv, L. & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science, 21(4)*, e12593. https://doi.org/10.1111/desc.12593

Romberg, A. & Saffran, J. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Review of Cognitive Science,* 1*(6)*, 906-914.

Romero, J. (1995). *Gestural Organization in Spanish: An Experimental Study of Spirantization and Aspiration*. PhD Dissertation, University of Connecticut.

Saffran, J., Aslin, R. & Newport, E. (1996a). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Saffran, J., Newport, E. & Aslin, R. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory & Language, 35*(4), 606-621.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*, 101–105.

Selkirk, Elisabeth. (1995). Sentence prosody: intonation, stress and phrasing. In *The Handbook*

of Phonological Theory, J. Goldsmith (ed.). London: Blackwell, 550–569.

Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavioral Research Methods, 49*(2), 418-432.

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition*, 177, 198-213.

Spencer, M., Kaschak, M., Jones, J., & Lonigan, C. (2015). Statistical learning is related to early literacy-related skills. *Reading and Writing*, 28, 467-490.

Sugahara, M., & Turk, A. (2009). Durational correlates of English sublexical constituent structure. *Phonology, 26*, 477-524.

Tamminen, J., & Gaskell, M.G. (2008). Newly learned spoken words show long-term lexical competition effects. *Quarterly Journal of Experimental Psychology, 61*(3), 361-371.

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.

Thiessen, E. D., & Saffran, J. R. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. Language learning and development, 3(1), 73-100.

Turk, A. & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics, 28*, 397-440.

Tyler, M.D. & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *The Journal of the Acoustical Society of America, 126*(1), 367-376.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (2013). *Comprehensive Test of Phonological Processing—Second Edition (CTOPP-2)*. Austin, TX: Pro-Ed.

Warker, J. A. (2013). Investigating the retention and time course of phonotactic constraint

learning from production experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(1), 96-109.

Warner, N. & Tucker, B. (2011). Phonetic variability of stops and flaps in spontaneous and careful speech. *Journal of the Acoustical Society of America, 130*(3), 1606-1617.

Wechsler, D. (2011). *Wechsler Abbreviated Scale of Intelligence - Second Edition (WASI-II)*. San Antonio, TX: Psychological Corporation.

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

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America,* 91, 1707-1717.

Wiig, E., Semel, E., & Secord, W. (2013). *Clinical Evaluation of Language Fundamentals—Fifth Edition (CELF-5)*. Bloomington, MN: Pearson.

Ziegler, J., Pech-Georgel, C., George, F., & Lorenzi, C. (2011). Noise on, voicing off: Speech perception deficits in children with Specific Language Impairment. *Journal of Experimental Child Psychology, 110*, 362-372.