10-21-2023

Of words and whistles: Statistical learning operates similarly for identical sounds perceived as speech and non-speechOf words and whistles: Statistical learning operates similarly for identical sounds perceived as speech and non-speech

Sierra J. Sweet

Stephen C. Van Hedger

Laura J. Batterink
lbatter@uwo.ca

Of words and whistles: Statistical learning operates similarly for identical sounds perceived as speech and non-speech

Sierra J. Sweet [1] (ssweet4@uwo.ca)

Stephen C. Van Hedger [1 2 3] (svanhedg@uwo.ca)

Laura J. Batterink [1 2] * (lbatter@uwo.ca)

[1] Department of Psychology, Western University, London, ON Canada

[2] Western Institute for Neuroscience, Western University, London, ON Canada

[3] Department of Psychology, Huron University College, London, ON Canada

* Corresponding author: Laura J. Batterink (lbatter@uwo.ca); (519) 661-2111 x85409; 1151 Richmond St. WIRB 6124, London, ON Canada, N6A 5B7

Abstract

Statistical learning is an ability that allows individuals to effortlessly extract patterns from the environment, such as sound patterns in speech. Some prior evidence suggests that statistical learning operates more robustly for speech compared to non-speech stimuli, supporting the idea that humans are predisposed to learn language. However, any apparent statistical learning advantage for speech could be driven by signal acoustics, rather than the subjective perception *per se* of sounds as speech. To resolve this issue, the current study assessed whether there is a statistical learning advantage for ambiguous sounds that are subjectively perceived as speech-like compared to the same sounds perceived as non-speech, thereby controlling for acoustic features. We first induced participants to perceive sine-wave speech (SWS)—a degraded form of speech not immediately perceptible as speech—as either speech or non-speech. After this induction phase, participants were exposed to a continuous stream of repeating trisyllabic nonsense words, composed of SWS syllables, and then completed an explicit familiarity rating task and an implicit target detection task to assess learning. Critically, participants showed robust and equivalent performance on both measures, regardless of their subjective speech perception. In contrast, participants who perceived the SWS syllables as more speech-like showed better detection of individual syllables embedded in speech streams. These results suggest that speech perception facilitates processing of individual sounds, but not the ability to extract patterns across sounds. Our findings suggest that statistical learning is not influenced by the degree of perceived linguistic relevance of sounds, and that it may be conceptualized largely as an automatic, stimulus-driven mechanism.

Keywords: statistical learning, speech, sine-wave speech, auditory perception

1. Introduction

Statistical learning, our ability to become sensitive to patterns in the environment, has provided an important mechanistic explanation for language acquisition since its initial documentation in the context of speech segmentation (Saffran et al., 1996a). In this study, infants were presented with a continuous stream of trisyllabic nonsense words, with no pauses or other acoustic cues to mark word boundaries. Thus, the probabilities of syllables co-occurring with one another provided the only indication of where individual words started and ended within the stream. After listening to the stream, infants were able to successfully discriminate between words and foil items through their looking time behaviour, providing evidence that they had extracted the statistical information in the stream to discover the embedded words.

Since this seminal study, subsequent research has shown that statistical learning is present across many domains outside of language (e.g., Conway & Christiansen, 2005; Fiser & Aslin, 2001**;** Saffran et al., 1999**;** Van Hedger et al., 2022). In one such study, conducted by Saffran and colleagues (1999), participants were exposed to a stream of six “tone words,” each of which consisted of a sequence of three pure tones. On a subsequent two-alternative forced-choice recognition task, participants succeeded in discriminating between tone words and foil sequences, providing a clear demonstration that statistical learning also operates across non-linguistic auditory stimuli – that is, auditory stimuli that lack a clear communicative purpose. Subsequent research has found that listeners can also extract patterns embedded in non-linguistic noises (Gebhart et al., 2009), everyday environmental sounds (Siegelman et al., 2018), tactile sequences (Conway & Christiansen, 2005), visual stimuli (e.g., Bulf et al., 2011; Kirkham et al., 2002; Fiser & Aslin, 2001), and multimodal contexts (Mitchel et al., 2014; Seitz et al., 2007). Further, statistical learning is present not only in infants but also in older children and adults (e.g., Moreau et al., 2022; Raviv & Arnon, 2018; Saffran et al., 1996b, 1997), as well as in nonhuman animals, including dogs (Boros et al., 2021) and cotton-top taramins (Hauser et al., 2001). These observations have led to a general consensus that statistical learning is not a “special” language-specific mechanism, but is domain-general in that it is present across modalities, domains, and even species (Aslin, 2017).

However, while statistical learning may be considered domain-general in that it is present in many learning contexts, it shows important differences depending on stimulus modality and learning domains, suggesting that it may not be a truly unitary mechanism (Frost et al., 2015; Frost et al., 2019). For example, an early study found an advantage for statistical learning of non-linguistic tones, as compared to tactile and visual stimuli, which persisted even after controlling for low-level perceptual differences between stimuli (Conway and Christiansen, 2005). Another study reported that changes in presentation rate have opposite effects on auditory and visual statistical learning: auditory statistical learning benefits from faster presentation rates, whereas visual statistical learning benefits from slower rates (Emberson et al., 2011). In addition, different types of statistical learning follow different developmental trajectories; statistical learning for speech sounds is stable from childhood into adulthood; in contrast, statistical learning improves with age for visual stimuli and non-linguistic tones (Arciuli &

Simpson, 2011; Moreau et al., 2022; Raviv & Arnon, 2018; Schlichting et al., 2017; Shufaniya & Arnon, 2018; for review, Forest et al., 2023).

These findings, which indicate that statistical learning is not equivalent across modalities, are not easily accommodated within frameworks that treat statistical learning as a single unitary mechanism. Further evidence against a unitary view of statistical learning comes from low interindividual correlations in statistical learning performance across modalities and stimulus materials (Siegelman & Frost, 2015; Siegelman et al., 2017). While an individual's statistical learning performance within a given domain is relatively stable, as assessed by test-retest reliability, performance on one task does not predict performance on a parallel tasks in a different domain (e.g. syllables to visual shapes; Siegelman & Frost, 2015). Taken together, these results suggest that there are nonoverlapping mechanisms supporting statistical learning abilities in different domains, supporting a "pluralist" view of statistical learning (Frost et al., 2015; Frost et al., 2019). According to this viewpoint, statistical learning is supported not only by domain-general mechanisms (e.g. Schapiro et al., 2014; Covington et al., 2018; Conway, 2020; Batterink et al., 2019), but also by modality-specific mechanisms that are united by similar computational principles. These modality-specific mechanisms operate within distinct networks and are governed by different constraints, depending on task domain and modality (Frost et al., 2015, Frost et al., 2019; Conway, 2020).

1.1. Is speech a privileged target for statistical learning?

The consensus that there are important differences in statistical learning as a function of learning domain raises a more specific question of whether statistical learning operates differently—and perhaps more robustly—for speech than non-speech.

Human infants prefer to listen to speech compared to other auditory stimuli (Shultz & Vouloumanos, 2010), and neuroimaging studies in adults have found greater activation in left auditory cortex for speech compared to other sounds (Binder et al., 2000; Narain et al., 2003; Parviainen et al., 2005; Scott et al., 2000; Vouloumanos et al., 2001). These results are in line with the general idea that speech is “special,” engaging unique neural and cognitive mechanisms not engaged by other auditory stimuli (Belin et al., 2000; Liberman, 1982; Marno et al., 2015; Moore, 2000).

Infant studies of artificial grammar rule learning also support this notion, suggesting that babies more readily extract simple grammar rules (e.g., “AAB” or “ABB” rules) from speech than from non-speech auditory stimuli, such as tones or animal sounds (Dawson & Gerken, 2009; Marcus et al., 2007). A number of theoretical hypotheses (which are not mutually exclusive) have been proposed to account for this speech advantage in rule learning, including that speech (1) better captures and holds infants’ attention (Schultz & Vouloumanos, 2010; Vouloumanos & Werker, 2004), (2) represents a communicative signal (Rabagliati et al., 2012; Ferguson & Lew-Williams, 2016), (3) is more familiar than other signals to infants, which facilitates learning (Saffran et al., 2007; Thiessen, 2012), and/or (4) may be processed by specific mechanisms that have been tuned to speech as humans evolved the capacity for language (Rabagliati et al., 2012; Marcus & Rabagliati, 2008, as cited in Ferguson & Lew-Williams, 2016). By extension, speech could also represent a privileged target for the statistical learning of embedded units in continuous sound sequences, in infants and adults alike.

Current evidence on whether there is indeed a statistical learning advantage for speech sounds is conflicting. A recent study by Ordin and colleagues (2021) supports the idea that there is a speech advantage in statistical learning. Participants were presented with embedded triplet sequences that were fully linguistic in nature (made up of natural syllables), semi-linguistic (made up of syllables that contained atypical acoustic cues), and non-linguistic (made up of environmental sounds such as animal noises and footsteps), and then asked to make old/new judgments for triplets from the sequences and foils. Performance was highest in the syllable condition compared to the semi-linguistic and non-linguistic conditions, providing support for a speech advantage for statistical learning. This result also converges with rule learning studies in infants, which have found a general advantage for speech stimuli over non-speech stimuli, as described above (e.g., Dawson & Gerken, 2009; Marcus et al., 2007).

However, not all studies point to a clear linguistic advantage for statistical learning. In the previously described “tone words” study by Saffran and colleagues (1999), both age groups successfully segmented the tone stream, and no significant differences were found between their performance on the tone version and the syllable version of the task from a previous study (Saffran et al., 1996b). Similarly, another study by Saffran (2002) presented adults and children with linguistic or non-linguistic auditory “sentences,” made up of nonsense words for the linguistic group (e.g. kiff flor lum dupp) and sequences of sounds such as bells, chimes, and drums for the non-linguistic group. Both groups learned successfully and again, no significant differences were found between conditions. Finally, a more recent study by Siegelman and colleagues (2018) compared statistical learning of syllables and everyday environmental sounds. Overall performance was similar between the two conditions, again suggesting that statistical learning occurs with similar efficacy for speech and non-speech sounds.

Yet, even in situations where overall learning is comparable for linguistic and non-linguistic items, there is evidence that linguistic items still might exhibit distinct patterns of learning. For example, more nuanced analyses of the Siegelman and colleagues (2018) data revealed that individual test items in the syllable condition showed much lower internal consistently than in the sound condition. Additional experiments indicated that participants' performance was influenced by the degree to which test items corresponded to the phonotactics of their own native language of Hebrew (see also Elazar et al., 2022). These results suggest that learners' prior knowledge and expectations may critically impact statistical learning of linguistically-relevant speech sounds, an effect that is less pronounced for non-linguistic sounds (though see Van Hedger et al., 2022 for evidence of effects of prior knowledge on statistical learning of instrument notes). Thus, even in the absence of overall performance differences, there may be qualitative differences in how statistical learning operates for speech versus non-speech sounds, particularly with respect to how learning interacts with other cognitive factors.

1.2. Differences between speech and non-speech sounds

Part of the difficulty in assessing whether there may be a statistical learning advantage for speech is that speech sounds and non-speech sounds, such as tones and environmental noises, differ in many ways. Previous learning studies comparing speech and non-speech have used different types of artificial languages, different syllable inventories, and many different types of non-linguistic sounds (e.g. Marcus et al., 2007; Ordin et al., 2021; Saffran et al., 1999; Saffran, 2002; Siegelman et al., 2018). Thus, conflicting results across studies could—in principle—be at least partially attributable to surface features of the learning materials. For example, speech sounds and other natural auditory stimuli such as musical instruments and everyday object sounds differ in fundamental frequency, timbre, aperiodicity, spectral variability, spectral envelope, and temporal envelope (Ogg & Slevc, 2019). Any number of these low-level acoustic features that differ between speech and non-linguistic stimuli may influence perception, ease of encoding, and consequently statistical learning performance. In other words, statistical learning differences between speech and non-speech—when observed—could reflect signal-driven differences in lower-level processes, such as the perception of individual items, rather than statistical learning *per se*.

A study by Thiessen (2012) highlights the importance of considering acoustic features when comparing statistical learning of speech versus non-speech sounds. The authors of this study reasoned that speech contains more redundant cues to an abstract rule than are typically available in non-linguistic stimuli, and that such redundancy may facilitate rule learning. For example, a string such as “ga ti ga” instantiates the “ABA” rule at multiple levels: at the syllable level, at the individual phoneme level (both the initial consonant and final vowel differentiate the A and B elements) and at the level of phonetic features (e.g., voicing). To test the importance of redundancy, the authors presented infants with syllable sequences that contained reduced redundancy, in which only the vowels, rather than both vowels and consonants, signaled the underlying rule (e.g. “ba bi ba” rather than “ga ti ga”). When redundancy was reduced, infants’ rule learning was impaired, suggesting that speech may allow for easier learning than non-linguistic stimuli at least in part because of the redundant information in the acoustic signal. These results underscore the importance of accounting for acoustic differences in comparisons of statistical learning between speech and non-speech stimuli.

In addition to their acoustic differences, speech sounds also differ from non-speech sounds in terms of their *subjective value* or *perceived relevance* to the listener. In contrast to tones or environmental noises, speech sounds are a linguistically relevant signal and serve a critical communicative purpose. This communicative value could in part explain why speech captures infants' attention to a greater degree than non-speech (e.g., Vouloumanos & Werker, 2004, 2007; Vouloumanos et al., 2010), or why auditory-relevant regions within the left temporal lobe are more strongly activated for speech than non-speech (Belin et al., 2000; Binder et al., 2000; Dick et al., 2007; Scott et al., 2000), although here too acoustic differences cannot be ruled out. To our knowledge, no previous studies have directly examined whether the communicative value of speech *per se* may play a role in potential statistical learning differences between speech and non-speech sounds.

In the current study, we tested the hypothesis that speech may serve as a privileged target for statistical learning due to its *subjective value* as a communicative signal, over and above any effects of acoustic differences between speech and non-speech. To address this hypothesis, we leveraged "sine-wave speech" (SWS), a manipulation that allows for comparing the processing of identical acoustic stimuli that may be perceived from highly speech-like to un-speechlike. SWS is a degraded form of natural speech consisting of time-varying sine waves modelling formant frequencies, with fewer sine waves corresponding to greater degradation of the signal (Remez et al., 1981). This degraded audio retains the phonetic properties of the original speech, but typically fails to be perceived as phonetic by naïve listeners, who may experience it as a sequence of whistles, chirps, and other types of “science fiction” sounds. SWS lacks many of the acoustic features that make speech sound natural, such as a fundamental frequency. However, it can still be perceived as speech if instructions to attend to the speech-like qualities of the stimuli, or information about its true nature, are given. For example, participants may suddenly perceive SWS as speech if they are played the intact, original audio immediately prior to the SWS version. Notably, once participants are induced into perceiving the SWS as speech, there is no known method to revert them back into hearing it as non-speech (Silva & Bellini-Leite, 2020). SWS thus provides a tool for manipulating listeners’ subjective, top-down perception of a signal as speech versus non-speech, while holding the physical stimuli constant. Essentially, this approach can be used to isolate speech-specific perceptual effects on statistical learning, independent of any acoustic differences.

1.3. The Current Study

The aim of the current experiment was to investigate whether statistical learning operates differently for sounds perceived as more speech-like compared to sounds perceived as non-speech in the absence of acoustic differences between stimuli. Participants initially completed an induction task, in which we attempted to induce them to perceive SWS syllables as either speech or non-speech sounds. They were then exposed to a continuous stream of repeating trisyllabic “words” composed of SWS syllables, and then completed two behavioural tasks to measure their statistical learning of the words: (1) an explicit familiarity rating task, in which participants rated their familiarity with the original words and two types of foil items and (2) a target detection task, which requires participants to make speeded responses to embedded syllables within continuous speech streams. This task does not require the conscious retrieval of previously learned information, providing an implicit measure of learning (Batterink et al., 2015). Finally, to determine each participants' subjective perception of the SWS, participants indicated on a 1-10 scale how speech-like they perceived the stimuli to be, and then transcribed SWS syllables and full SWS sentences.

As described previously, both low-level acoustic differences as well as high-level differences in perceived linguistic relevance could contribute to differences in statistical learning for speech versus non-speech sounds. Our experimental design allows us to isolate the role of subjective speech perception in statistical learning, independently of acoustic factors. If the subjective perception of sounds as linguistically relevant is an important factor for statistical learning, we would expect that learners who perceive the ambiguous SWS stimuli as speech-like to a greater degree to show better statistical learning performance on both measures. In contrast, if the primary factor driving differences in statistical learning of speech versus non-speech is the acoustic signal, we would expect no relationship between statistical learning performance and listeners' perception of the SWS stimuli, given that the stimuli themselves are identical. As we were interested in both positive and null findings, all analyses were substantiated with a Bayesian approach.

2. Method

2.1. Participants

A total of 200 participants were recruited from online participant recruitment platforms Prolific (n = 65; Palan & Schitter, 2018) and Amazon Mechanical Turk through CloudResearch (n = 135; Litman et al., 2017). Amazon Mechanical Turk participants were initially recruited; however, because a substantial proportion failed the study's attention check (as described in detail later), we recruited a second group of participants from Prolific in hopes of obtaining participants who would perform better on this attention check. All Amazon Mechanical Turk recruited participants were CloudResearch-approved, indicating that they had been screened and shown proof that they engage in tasks in an attentive manner. All Prolific participants had approval rates between 90-100%, indicating that a high percentage of their submissions for other research studies had been approved by the researchers. All participants reported English as their primary language, were above 17 years old, and had normal or corrected-to-normal hearing. Of the 200 participants, 100 were assigned to the speech induction condition, while the remaining 100 were assigned to the non-speech induction. Participants were financially compensated for their time.

Of the 200 participants, a total of 73 participants were excluded from analysis; 43 were excluded due to failing to pass both attention checks embedded in the exposure stream (as described in greater detail later); 23 because their data failed to save to our servers; and 3 due to making no responses during the target detection task. Finally, 1 participant was excluded due to not having normal or corrected-to-normal hearing, and 3 participants were excluded due to failing to meet the inclusion criteria of having English as their primary language, based off their answers to the post-study survey. Thus, final analyses comprise data from 71 participants in the speech induction (SI) condition (mean age = 40.2 y; SD = 11.8 y; 37 men; 34 women), and 56 participants in the non-speech induction (NSI) condition (mean age = 39.3 y; SD = 12.0 y; 29 men; 27 women).

2.2. Stimuli

The experimental stimuli consisted of 12 syllables recorded by a male native English speaker, taken from Batterink and Paller (2019), in addition to 24 corresponding SWS manipulated forms of these syllables, comprised of single-sine wave (highly degraded) and three-sine wave (moderately degraded) versions of each of the original syllables. Each syllable sound file was 300 ms. Manipulated forms of the syllables were created in Praat (Boersma & Weenick, 2022) using a script by Darwin (2003). The unmanipulated (original) forms and single-sine wave (highly degraded) forms of the syllables were used only as primes in the induction task. The three-sine wave (moderately degraded) forms comprised the key experimental stimuli that were used throughout all statistical learning tasks, as well as the syllable transcription task.

The 12 three-sine wave syllables were combined to create 4 trisyllabic nonsense words (e.g. *tafuko, rigimi, rupuni, fitisu*). To form the continuous artificial speech stream, these trisyllabic nonsense words were concatenated pseudorandomly, without pauses between words, with the constraint that the same word never occurred consecutively. Thus, the transitional probabilities of neighbouring syllables were 1.0 within a word, and 0.33 across word boundaries. The stream consisted of 600 syllables (200 words) presented at a rate of 300 ms per syllable (i.e. 3.3 Hz), with each of the 4 words repeated 50 times, for a total duration of 3 minutes. To control for potential syllable-specific idiosyncrasies, the syllables in a given word were each assigned to the first, second, and third position across three conditions, counterbalanced across participants (Language A: *tafuko, rigimi, rupuni, fitisu*; Language B: *fukota, gimiri, puniru, tisufi*; Language C: *kotafu, mirigi, nirupu, sufiti*). The experimental script was programmed in jsPsych (de Leeuw et al., 2023).

2.3. Procedure

All tasks were performed online on the participants' own laptops or personal computers. To minimize distractions during the study, participants were asked to complete the tasks in a quiet listening environment and to use headphones for the entire duration of the session. Each session began with a volume adjustment task during which participants listened to a thirty-second noise and adjusted their sound volume to a comfortable level.

The experimental procedure is summarized in Figure 1, and consisted of four main phases, as described below. Participants completed one of two different versions of the induction task depending on whether they were assigned to the SI or NSI condition. All other tasks, including the key SWS stimuli, were identical between groups.

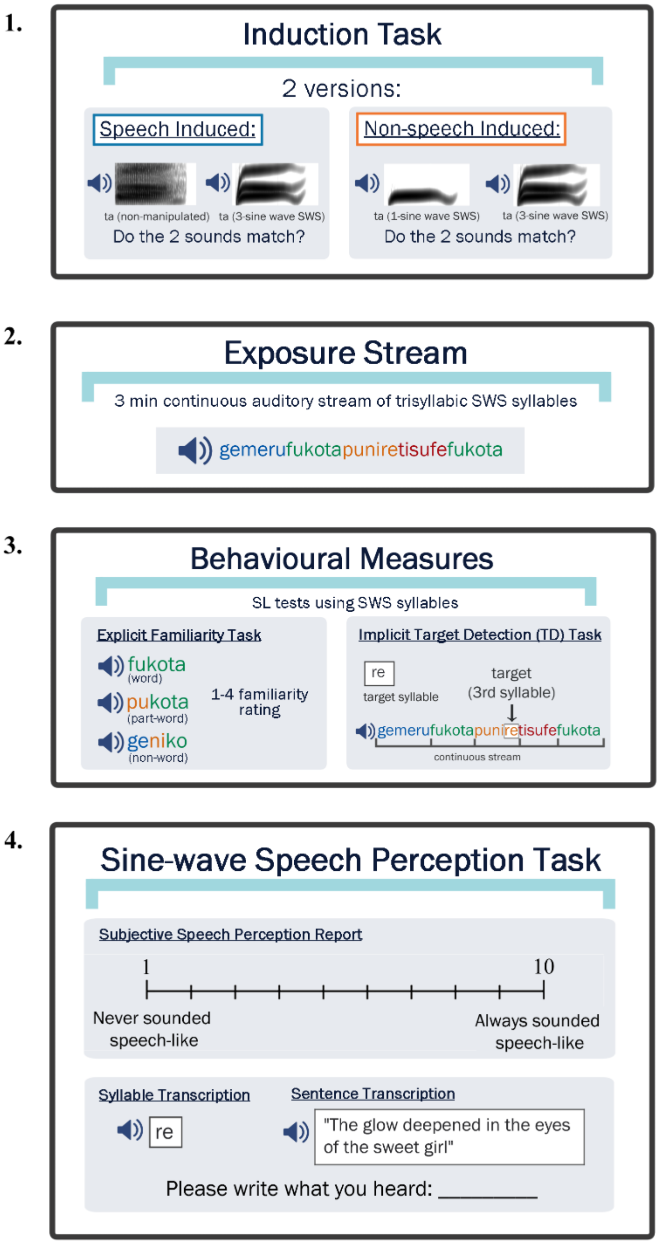


Figure 1. A summary of the experimental procedure. The induction task in the speech induced condition consisted of judging whether pairs of intact syllables and moderately degraded syllables matched. The induction task in the non-speech induced condition consisted of judging whether pairs of moderately degraded and heavily degraded syllables matched. Participants were exposed to 3 minutes of repeating nonsense words composed of the key SWS syllables. To measure learning, participants then completed a familiarity rating task, in which they rated the familiarity of words and foils, and a target detection task, in which they responded each time they detected a target syllable in a continuous stream consisting of the nonsense words. Finally, for each of

the 12 key SWS syllables, participants were asked to indicate how speech-like they thought they were, and then transcribed the SWS syllables and sentences to the best of their ability. Task order was identical for all participants.

2.3.1. Induction Task

This task was designed to induce participants to perceive the key SWS stimuli as either speech (SI condition) or as non-speech (NSI condition). In this task, participants were presented with “matched pairs” of syllables and instructed to intentionally learn the syllable pairings. The SI participants were told that they would be listening to speech syllables, and that each syllable would be followed by a distorted version of itself. They were then presented with syllable pairs comprised of the intact, non-manipulated version of each syllable (e.g. “fu”) followed by the target SWS version of the same syllable (e.g. the three-sine-wave version of “fu”), in order to draw their attention to the speech-like qualities of the SWS syllables. In contrast, the NSI participants were told that they would be listening to robotic noises artificially generated by a computer. The NSI participants were then presented with syllable pairs consisting of the highly degraded version of each syllable (e.g. the single-sine wave version of “fu”) followed by the target SWS version.

The task was made up of an initial training phase, followed by a test phase. In the training phase, participants were simply presented with two repetitions of each of the 12 pairs (24 total trials) and were instructed to pay careful attention as they would be tested on the pairs later. Next, participants completed 40 test trials, comprised of 36 correctly paired syllables and 4 mismatched pairs. On each test trial, participants were asked to judge whether the two sounds made up a correctly matched pair by pressing one of two corresponding keys.

2.3.2. Exposure Stream

Next, participants were presented with the three-minute continuous stream of nonsense words, made up of the same key SWS syllables for both induction groups. They were instructed to pay attention to the stream, and were told they may be tested on their knowledge of the stream later in the study. To ensure participant engagement in the online testing environment, two attention checks were embedded within the exposure stream, consisting of 4 s pauses inserted randomly at two of nine preselected times in the stream. Prior to beginning the task, participants were instructed to listen for pauses and to press the spacebar key within 4 s whenever they heard a pause. Failure to detect both pauses resulted in participant exclusion from subsequent analyses.

2.3.3. Statistical Learning Tasks

Next, participants completed two behavioural tests of statistical learning, in the order indicated below.

2.3.3.1. **Familiarity Rating Task**

This task is designed to assess explicit memory of the nonsense words (e.g. Batterink & Paller, 2017, 2019). On each trial, participants listened to a syllable triplet made of the key SWS syllables, and rated how familiar it sounded to them on a scale from 1 (*very unfamiliar)* to 4 (*very familiar*). A total of 12 trials were presented, with 4 trials consisting of words from the exposure stream (e.g. *tafuko)*, 4 trials consisting of part-words (i.e. a syllable pair from a word in the exposure stream combined with an additional syllable from a different word, e.g. *rufuko*), and 4 trials consisting of non-words (syllables from the stream that had never occurred together, e.g. *rupufu*).

Evidence of explicit memory for the words would be provided by higher ratings to words, followed by part-words, with non-words rated as least familiar.

2.3.3.2. Target Detection Task

This task measures participants' response times to target syllables embedded within shortened versions of the speech stream, and can reveal statistical learning in the form of prediction effects, in the absence of explicit memory or intentional retrieval of the learned words (Batterink et al., 2015). On each trial, participants were presented with a target SWS syllable; they were allowed to replay this target syllable as many times as they wished. They then listened to a shortened version of the exposure stream (~14.5 s), containing the four trisyllabic nonsense words concatenated together four times each in pseudorandom order (48 syllables total), in the same manner as the Exposure stream. Participants were instructed to press the spacebar each time they heard that target syllable as quickly and accurately as possible by pressing the spacebar.

Each of the 12 SWS syllables acted as a target three times overall, yielding a total of 36 streams. Across all streams, this yielded a total of 144 targets, 48 within each syllable position (1^{st}, 2^{nd}, 3^{rd}). Stream order was randomized for every participant. Successful learning of the speech stream would be reflected by faster reaction times to target syllables that occurred in the medial or final position of a trisyllabic word relative to syllables that occurred in the initial position, due to the opportunity to predict the target (Batterink et al., 2015, 2019; Batterink & Paller, 2017).

2.3.4. Speech Perception Task

This task was designed to examine participants' perception and comprehension of the key SWS stimuli, and contained three parts. As illustrated in Figure 1, this was always the final task in the experiment, in order to avoid suggesting the communicative nature of SWS to participants in the NSI group.

2.3.4.1. Overall Subjective Speech Perception Rating.

Participants were presented with an open-response textbox and asked to describe the sounds that they had heard in the study. Using a slider, they were then asked to rate the extent to which they had heard the SWS as speech-like, with the scale ranging from 1 (*I never heard the sounds as speech*) to 10 (*I always heard the sounds as speech*).

2.3.4.2. Syllable Transcription

Participants then listened to each of the 12 key SWS syllables one at a time and were asked whether they thought it sounded like speech (yes/no response). If a participant indicated that they heard a syllable as speech, they were then asked to transcribe the syllable to the best of their ability by typing their response into an open-response textbox.

2.3.4.3. Sentence Transcription

As a test of generalized SWS perception, participants listened to 10 SWS sentences from the Harvard sentences database (IEEE, 1969) and transcribed each one to the best of their ability. An example of one of the sentences is, "The glow deepened in the eyes of the sweet girl." Participants were instructed to spell each word as accurately as possible.

2.3.5. Survey

Finally, participants were redirected to a Qualtrics survey containing basic demographic questions about age, gender identity, and language fluency.

2.4. Statistical Analyses

For all t-tests, the Student's t-test was utilized unless the assumption of equal variances was violated. Welch's unequal variances t-tests were instead used whenever Levene's Test was significant.

Bayes Factors were calculated for each test, using the default prior provided by JASP. This prior uses a Cauchy distribution, centered around 0, with a width parameter of 0.707. The reported Bayes Factors (BF_{10}) represent how likely the alternative hypothesis is relative to the null hypothesis; values above 1 indicate evidence supporting the alternative hypothesis, whereas values below 1 provide evidence supporting the null hypothesis over the alternative hypothesis. As an example, a BF_{10} of 4 indicates that, given the data, the alternative hypothesis is four times likelier than the null hypothesis. In contrast, a BF_{10} of 0.25 would indicate that the alternative hypothesis is one-fourth as likely as the null hypothesis. Conventional means of interpreting the relative strength of Bayes Factors regard BF_{10} = 3-10 as moderate evidence, such that a BF_{10} of 4 suggests moderate evidence for the alternative hypothesis over the null hypothesis (Schmalz et al., 2023). Bayes Factors can also be reported using BF_{01}, the inverse of BF_{10}, which presents the likelihood of the null hypothesis relative to the alternative hypothesis. Thus, a BF_{01} of 4 indicates that the null hypothesis is four times likelier than the alternative hypothesis. BF_{10} values are reported for each test in this study; however, for any tests that result in null findings, BF_{01} is also be reported for ease of interpretation.

2.4.1. Induction Task

Each participant's accuracy on the matched pairs test was calculated. Additionally, as there were many more "match" trials than "mismatch" trials, we also computed d' scores as a bias-free measure of participants' sensitivity to the presence of a match. D' was computed as the difference between the z-transforms of participants' hit rate (i.e. the proportion of matched trials that they correctly identified as matching) and false alarm rate (the proportion of mismatched trials that they incorrectly identified as matching) in the task.

2.4.2. Statistical Learning Tasks

For all analyses of the statistical learning tasks, Greenhouse–Geisser corrections were reported for factors with more than two levels.

2.4.2.1. Familiarity Task

Average familiarity ratings were computed for each word category (Word, Partword, Nonword) and entered into a 2x3 mixed effects ANOVA with induction condition (speech induced, non-speech induced) as a between-subjects factor and word category (non-word, part-word, word) as a within-subjects factor.

Additionally, for subsequent correlational analyses, "familiarity rating scores" (Batterink & Paller, 2017, 2019) were calculated by subtracting the average of a participants' rating of partwords and nonwords from their average rating of a word.

Perfect sensitivity to words over foils on this measure would be a score of 3, with any positive value suggestive of learning, as this would reflect higher scores for words compared to both pseudo- and non-words.

2.4.2.2. Target Detection Task

Following the inclusion criteria of previous studies, responses that occurred within 1200 ms following target onset were considered valid hits (Batterink & Paller, 2017, 2019). All other responses were considered false alarms.

2.4.2.2.1. Detection Score

For each participant, we first calculated the number of targets that were correctly detected and the total number of false alarms. We then computed an overall "detection score," which represents a conservative estimate of a participant's sensitivity to the targets in the stream, computed as the overall number of hits divided by the overall number of false alarms (Number of Hits/Number of False Alarms). Given that the "target response" window (4 targets x 1200 ms = 4800 ms) for each stream was half the length of the "false alarm" windows (total stream length of 14400 ms – "target response" length of 4800 ms = 9600 ms), we reasoned that any score *greater* than 0.5 would provide evidence of above-chance detection performance (with 0.5 indicating that hits occurred half as frequently as false alarms, as would be expected if responses were distributed randomly across the stream, without regard for the actual target locations). In other words, a detection score of >0.5 would indicate that participant's responses were more likely to occur within a "target response" window than a "false alarm" window, providing evidence of target detection at above-chance levels.

2.4.2.2.2. Reaction Time

In addition to already-reported exclusions (see section 2.1.), 3 additional participants who only responded to initial targets were excluded from the RT analysis, as their mean response times could not be computed for second and third position targets. Furthermore, participants with a detection score of 0.5 or below were also excluded from this analysis (n = 32). We reasoned that if a participant is unable to detect the syllables at an above-chance level, any differences in their RTs cannot be considered a valid measure of statistical learning. To summarize, 35 additional participants were excluded from this analysis, yielding a final *n* of 92 participants. 52 of these participants completed the speech induction (mean age = 40.0 y, SD = 11.5 y; 28 men; 24 women), and the remaining 40 were from the NSI group (mean age = 39.7 y, SD = 13.1 y; 19 men; 21 women). For thorough reporting, a parallel analysis that also includes data from participants who scored below chance on detection can be found in Supplementary Materials (n = 124).

For each participant, mean RTs for detected targets were calculated for each target position (initial, medial, final). Mean RTs were then entered into a 2 x 3 repeated-measures ANOVA with induction group as the between-subject factor and target position (initial, medial, final) as the within-subject factor. In addition, to quantify statistical learning performance using a single metric while controlling for individual differences in baseline response times, a “RT prediction score” was computed by subtracting the average RT for the final syllable position from the average RT for the initial syllable position and dividing it by the average RT for the initial syllable position [$(RT_1-RT_3)/RT_1$; Batterink & Paller, 2019]. This calculation adjusts for potential differences in baseline RTs between individuals, allowing us to measure statistical learning across individuals with different RT baselines.

2.4.3. Speech Perception Tasks

2.4.3.1. Syllable Transcription

Scoring for this task was done by allocating 1 point for each syllable that was fully correctly transcribed (with alternative spellings such as “mee” or “me” designated as correct), and 0.5 points for each syllable that was partially correct, with either the consonant or vowel transcribed correctly (e.g. typing “mee” when the SWS syllable being played is “gee”). Average accuracy across the 12 total syllables in the task was then computed for each participant.

2.4.3.2. Sentence Transcription

Each SWS sentence contained 5 keywords (e.g. in the sentence “Pluck the bright rose without leaves” the keywords would be “pluck,” “bright,” “rose,” “without,” and “leaves”). While participants wrote out the entire sentence, their scores were calculated as the proportion of correctly transcribed keywords. Misspelled words were marked as incorrect.

3. Results

We first report the results from the induction task. Following this, we then characterize participants’ perception of the key SWS stimuli, as assessed through our three speech perception tasks (Figure 1). Although these speech perception tasks were completed at the end of the session, we report these results second, as they are needed to understand the subsequent statistical learning analyses. We then turn to our main set of results, which concerns performance on our two measures of statistical learning—the familiarity rating and the target detection tasks—and how performance on these tasks relates to perception of SWS stimuli.

3.1. Induction Task

Participants generally performed well on the matched pairs test, with an average accuracy rate of 90.7% (SD = 8.1%). Not surprisingly, given that they were presented with non-degraded syllable primes, speech induced (SI) participants outperformed non-speech induced (NSI) participants on this task (SI: mean = 94.9%; SD = 5.2%; NSI: mean = 85.4%; SD = 8.1%; $t(88.96) = -7.64$, $p < .001$, $d = -1.40$; $BF_{10} = 9.79 \times 10^9$).

The average d’ was 2.33 (SD = 1.05), with SI participants also outperforming NSI participants on this measure (SI: mean = 2.93; SD = 0.82; NSI: mean = 1.58; SD = 0.79; $t(125) = -9.39$, $p < .001$, $d = -1.68$; $BF_{10} = 1.27 \times 10^{13}$).

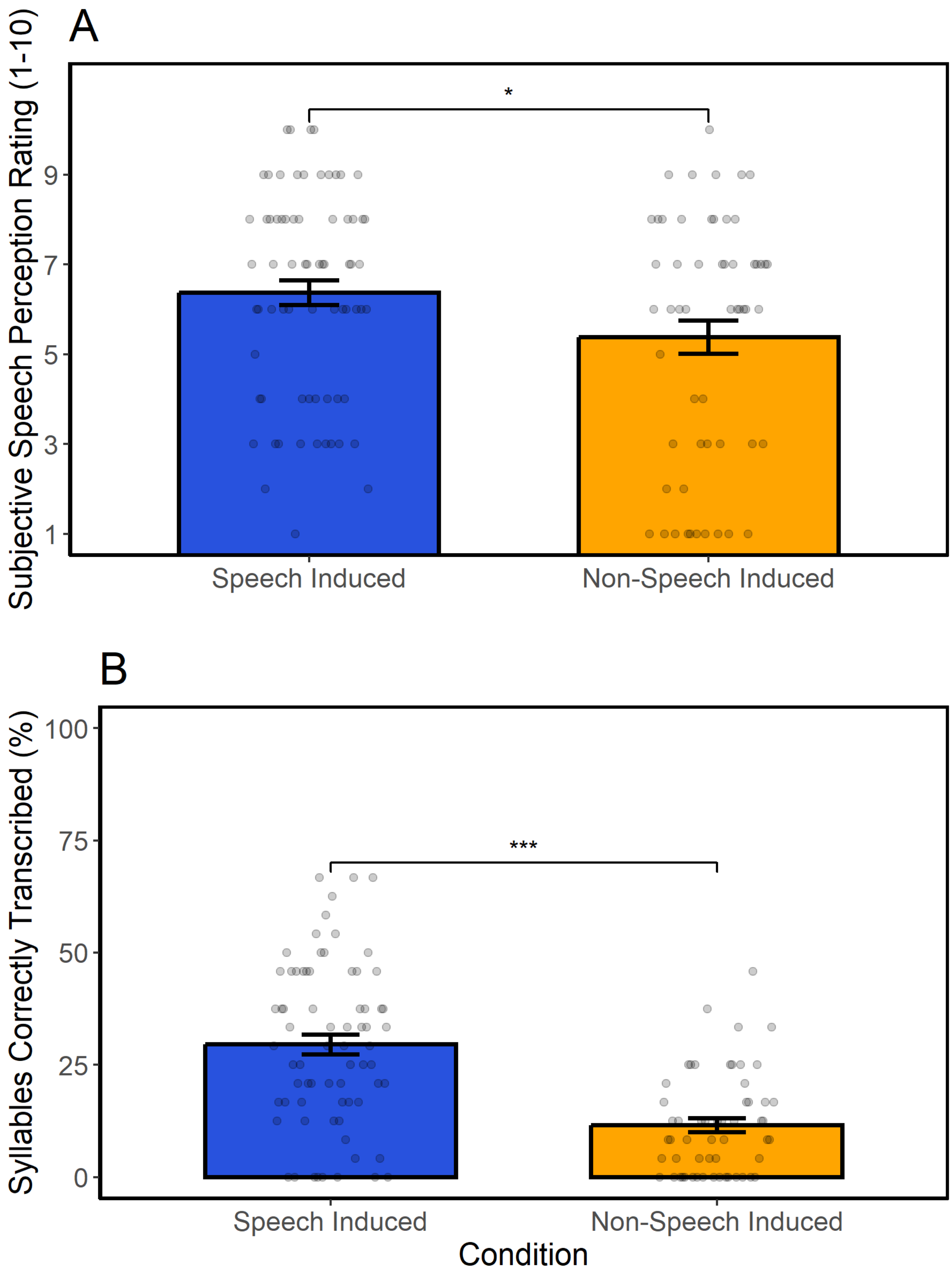
3.2. Speech Perception Tasks

3.2.1. Overall Subjective Speech Perception Rating

Reponses on the scale, ranging from 1 to 10, showed that SI participants (M = 6.37, SD = 2.32) rated the SWS as sounding significantly more speech-like overall than the NSI participants (M = 5.38, SD = 2.79), $t(106.71) = -2.14$, $p = .035$, $d = -0.39$; $BF_{10} = 1.62$. Nonetheless, there was considerable overlap in the scores, such that some NSI participants perceived the stimuli to sound more speech-like, while some SI participants perceived the stimuli to not sound very speech-like. The distribution of participant responses on the scale are presented in Figure 2A.

3.2.2. Syllable Transcription

As expected, participants in the SI group (M = 53.8%, SD = 28.3%) judged a significantly higher percentage of SWS syllables to be speech-like compared to the NSI participants (M = 35.7%, SD = 29.2%), $t(125) = -3.52$, $p < .001$, $d = -0.63$; $BF_{10} = 44.24$. Additionally, SI participants (M = 29.5%, SD = 18.3%) also correctly transcribed a significantly larger proportion of the 12 SWS syllables than the NSI participants (M = 11.5%, SD = 11.2%), $t(118.36) = -6.82$, $p < .001$, $d = -1.19$; $BF_{10} = 4.34 \times 10^6$ (see Figure 2B).



571

572 **Figure 2.** (A) The distribution of participant responses on the subjective speech
573 perception scale. The error bars represent the standard error of the mean. (B)

Participants' accuracy in syllable transcription task. The error bars represent the standard error of the mean. *p < .05; ***p < .001.

3.2.3. Sentence Transcription

Participants correctly transcribed 48.4% of the keywords in total (SD = 21.4%).

Somewhat unexpectedly, there was no significant difference in the keyword transcription accuracy between SI participants (M = 49.4%, SD = 22.2%) and NSI participants (M = 47.0%, SD = 20.5%), $t(125) = -0.64$, $p = .521$, $d = -0.12$; $BF_{10} = 0.23$ [$BF_{01} = 4.35$]. This suggests that the speech induction training on individual syllables did not generalize to novel sentences. However, across all participants, there was a significant positive correlation between accuracy on the syllable transcription task and sentence transcription task, $r(125) = 0.38$, $p < .001$; $BF_{10} = 1867.35$ (see Figure 3), suggesting that performance on these two tasks reflects a common ability.

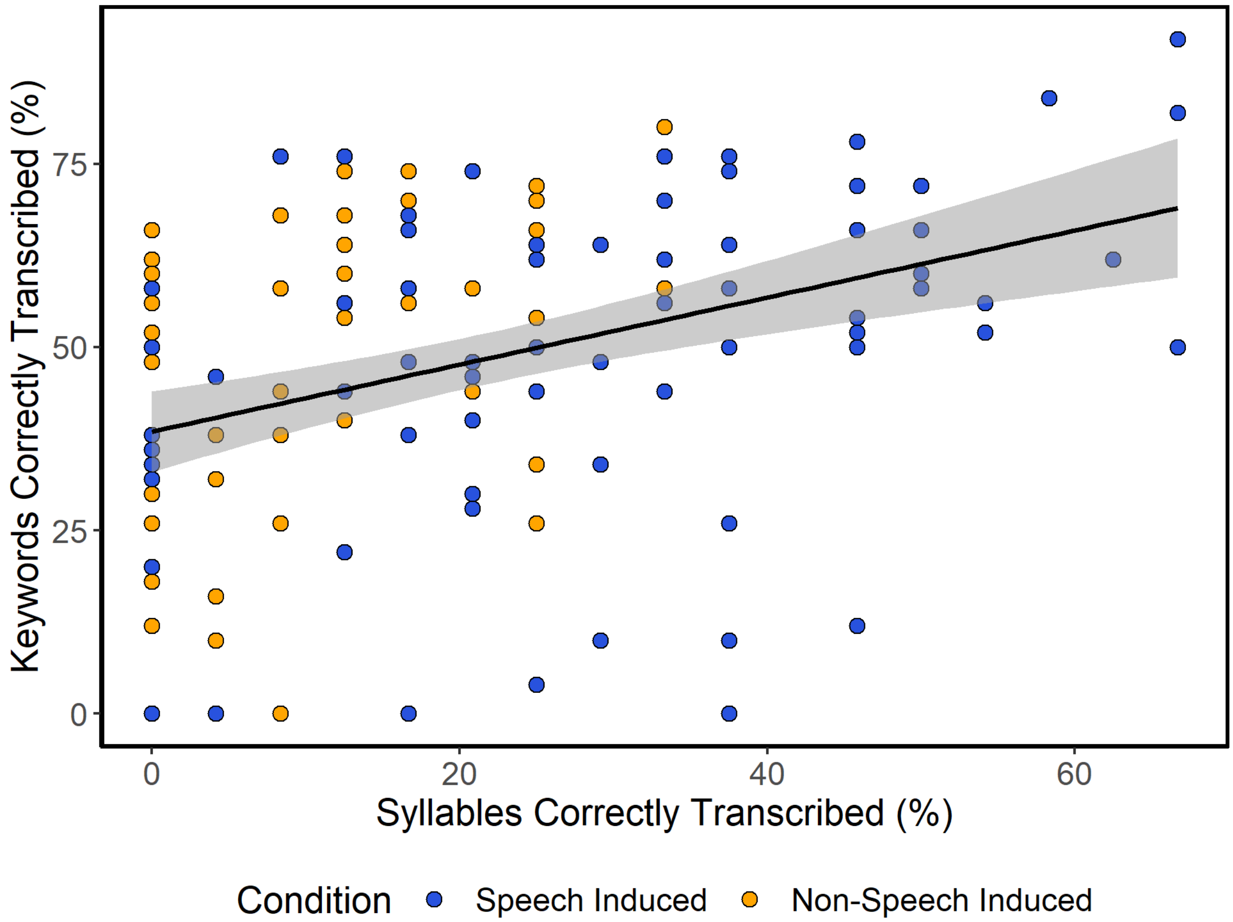


Figure 3. The correlation between the percentage of SWS sentences and the key SWS syllables that participants transcribed accurately ($r = 0.38$, $p < .001$).

3.3. Statistical Learning Tasks

As just described, while the two induction groups showed significant differences on self-reported subjective speech perception and on SWS syllable transcription accuracy, there was considerable overlap between the groups on these measures. In addition, there were no group differences on the sentence transcription task. These results indicate that our speech perception manipulation only partially altered participants' perception of the key SWS syllables, rather than producing a dramatic transformation of participants' percepts. Thus, as a further test of the relationship between statistical learning and speech perception, we examined correlations between participants' accuracy on the SWS syllable transcription task—taking this as a measure of speech perception—and their statistical learning performance. Hence, in the following section, for both our measures of statistical learning, we report (1) differences in performance between our two *a priori* defined groups and (2) correlations between accuracy on the syllable transcription task and statistical learning performance.

3.3.1. Familiarity Task

As expected, across both induction groups, words were rated as the most familiar, followed by part-words, with non-words rated as the least familiar, leading to a significant effect of word type, $F(1.98,248.18) = 18.00$, $p < .001$**,** $\eta^2p = 0.13$; $BF_{10} = 2.48 \times 10^5$ (see Figure 4A).

Supporting the hypothesis that statistical learning operates in a similar manner across stimuli perceived as linguistically-relevant and irrelevant, performance on the familiarity rating task was not significantly different between the two induction groups (Main Effect of Induction: $F(1,125) = 6.45 \times 10^{-3}$, $p = .936$**,** $\eta^2p = 5.16 \times 10^{-5}$; $BF_{10} = 0.22$ [$BF_{01} = 4.54$]; Word Type x Induction: $F(1.98,248.18) = 0.34$, $p = .714$**,** $\eta^2p = 0.0027$; $BF_{10} = 0.07$ [$BF_{01} = 14.3$]).

Further, there was no significant correlation between participants' syllable transcription accuracy and their familiarity rating scores, $r(125) = 0.09$, $p = .34$, with the Bayes Factor indicating moderate evidence (Schmalz et al., 2023) for the null hypothesis of no relation between these two measures ($BF_{10} = 0.18$ [$BF_{01} = 5.55$]; see Figure 4B). This result indicates that more accurate perception of the stimuli as syllables did not lead to better performance on the familiarity measure.

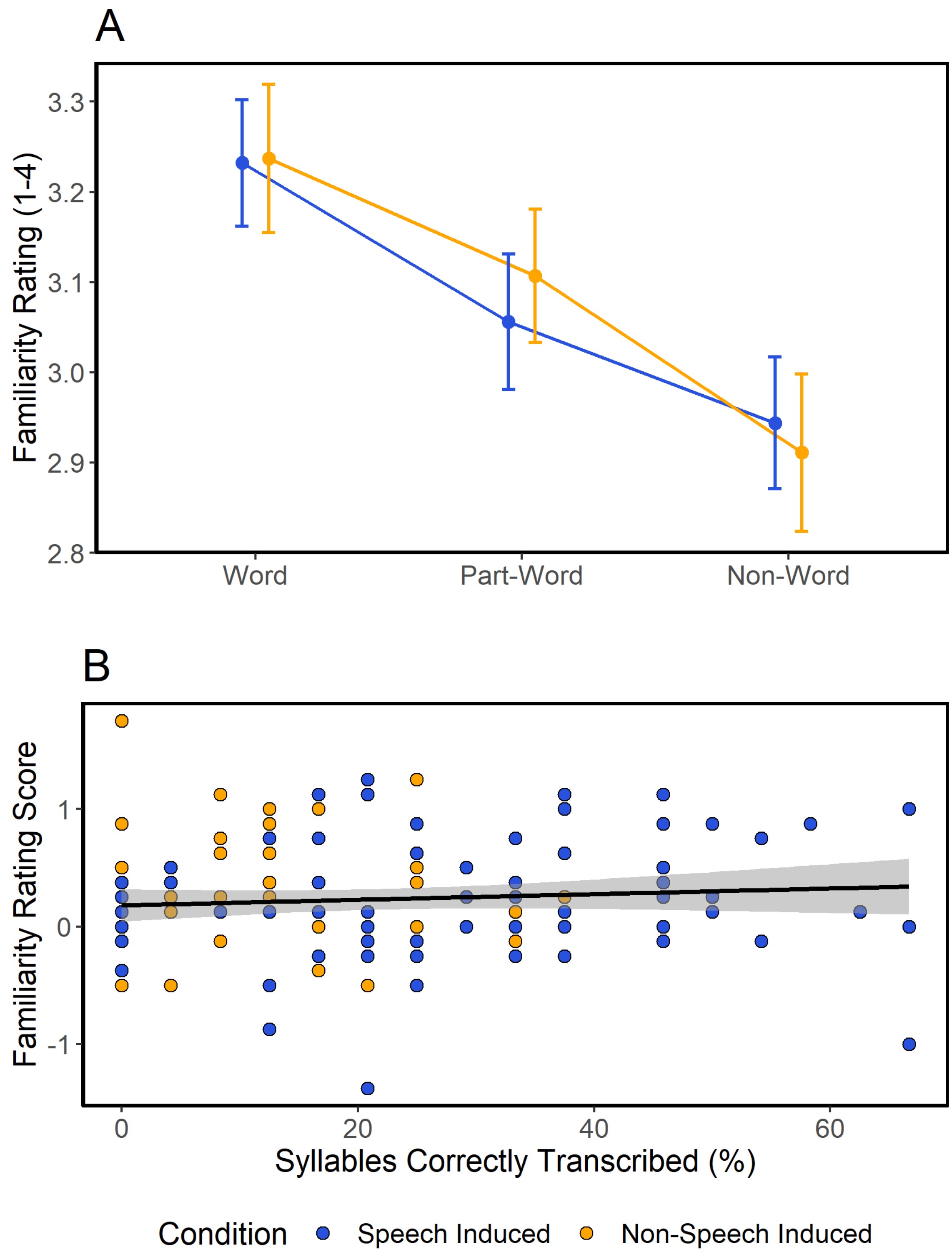


Figure 4. (A) Participants' ratings of triplet familiarity from the familiarity rating task. The error bars represent the standard error of the mean. (B) The correlation between

participants' familiarity rating score and the percentage of key SWS syllables that they transcribed accurately (r = 0.09, p = .34).

3.3.2. Target Detection Task

3.3.2.1. Overall Detection Rate

Participants correctly responded to 67.4% (SD = 20.0%) of the targets on average and made an average of 148.7 false alarms total (SD = 101.2). Accuracy rate was relatively low and false alarms were relatively high compared to previous versions of this task (e.g. Batterink et al., 2015; Batterink & Paller, 2017, 2019). This relatively poor performance may be attributed to the manipulated nature of the syllables, which made them more difficult to identify. Nonetheless, participants performed significantly above chance, as assessed by the detection score (M = 0.98, SD = 0.92; $t(126) = 5.91$, $p < .001$, $d = 0.52$; chance is 0.5 on this measure), with no significant difference in performance between the SI participants (M = 1.05, SD = 0.95) and NSI participants (M = 0.90, SD = 0.89), $t(125) = -0.93$, $p = .355$, $d = -0.17$; $BF_{10} = 0.28$ [$BF_{01} = 3.57$].

Interestingly, there was a significant positive correlation between the Detection Measure values and syllable transcription accuracy, $r(125) = 0.29$, $p = .001$; BF_{10} = 22.39, as presented in Figure 5. This result indicates that participants who more accurately perceived the stimuli as syllables were also better able to detect them in the continuous speech sequences.

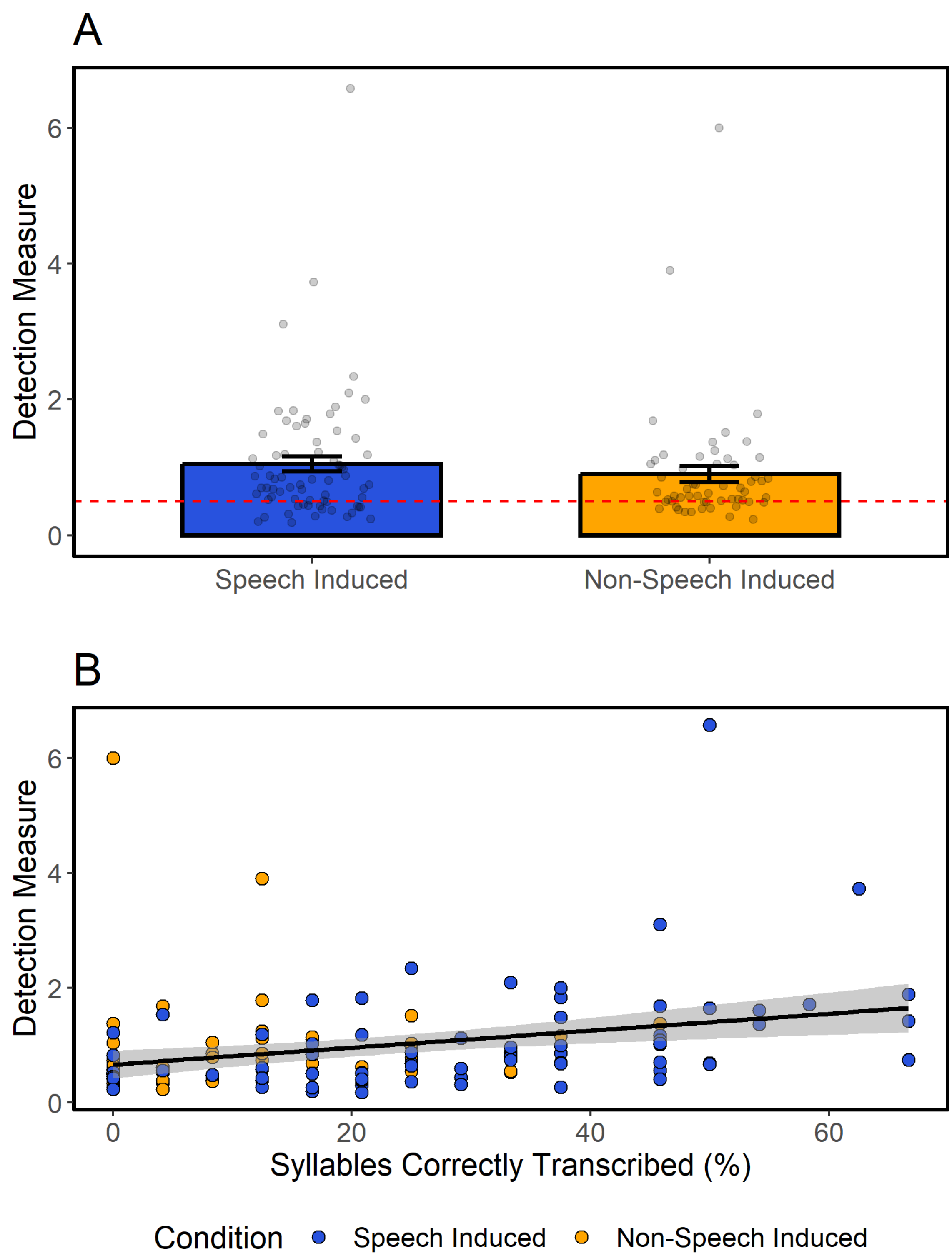


Figure 5. (A) Participants' detection score values on the target detection task (chance is 0.5). The error bars represent the standard error of the mean. (B) The correlation

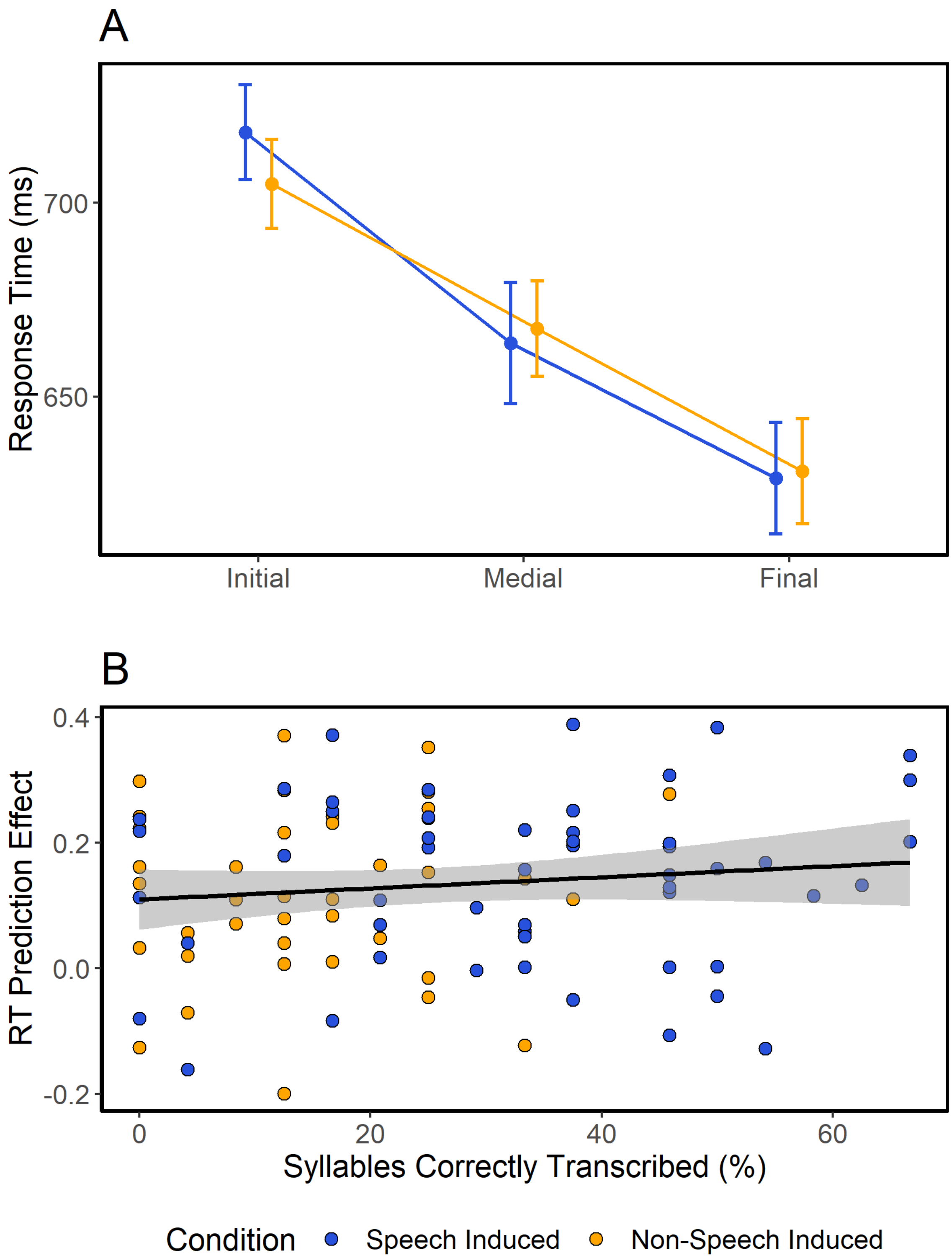
between participants' Detection Measure values on the target detection task and the percentage of key SWS syllables that they transcribed accurately (r = 0.29, p = .001).

3.3.2.2. Reaction Time

As expected, across both groups, RTs were the fastest for final-position syllables, second fastest for medial-position syllables, and slowest for initial-position syllables, as shown in Figure 6A, leading to a significant effect of syllable position, $F(1.68,150.76) = 61.69$, $p < .001$, $\eta^2p = 0.41$; $BF_{10} = 4.17 \times 10^{18}$. Notably, there was no significant difference in the RTs between induction groups, either overall or as a function of syllable position (Main Effect of Induction: $F(1,90) = 0.06$, $p = .802$, $\eta^2p = 7.00 \times 10^{-4}$; $BF_{10} = 0.24$ [$BF_{01} = 4.17$]; Position x Induction: $F(1.68,150.76) = 1.14$, $p = .315$, $\eta^2p = 0.01$; $BF_{10} = 0.19$ [$BF_{01} = 5.26$]).

Additionally, there was no significant correlation between RT prediction effect and syllable transcription accuracy, $r(90) = 0.12$, $p = .253$; $BF_{10} = 0.25$ [$BF_{01} = 4.00$], as shown in Figure 6B. This suggests more accurately perceiving the SWS stimuli as syllables did not lead to an enhanced ability to predict final position syllables. For a summary of the Bayes Factors for the study's statistical learning measures, see Table 1.

While the above analysis excludes participants who failed to detect syllables at above-chance levels, we also report results from the full sample (see Supplementary Materials). We note that the overall pattern of findings is largely similar between the two analyses.



672

673 **Figure 6.** (A) Participants' average reaction times for each of the syllable positions in
674 the target detection task. The error bars represent the standard error of the mean. (B)

The correlation between participants' RT prediction effect and the percentage of key SWS syllables that they transcribed accurately (r = 0.12, p = .253).

Table 1

Summary of Bayes Factor Results for Statistical Learning Performance

Task	BF_{01}	Strength of evidence in favour of null
Familiarity Task		
Main Effect of Induction	4.54	Moderate
Word Type x Induction	14.29	Strong
Correlation	5.55	Moderate
Target Detection Task		
Main Effect of Induction	4.17	Moderate
Position x Induction	5.26	Moderate
Correlation	4.00	Moderate

Note. Moderate evidence: BF_{01} = 3-10. Strong evidence: BF_{01} = 10-30. The null hypothesis here indicates no impact of speech perception on statistical learning performance.

4. Discussion

In the current study, we examined whether statistical learning occurs more robustly for sounds subjectively perceived as speech relative to those perceived as non-speech, independently of stimulus acoustics. The key novel aspect of the current study was the use of SWS to eliminate acoustic differences between stimuli perceived linguistically versus non-linguistically. Overall, we found that statistical learning operates similarly for stimuli, regardless of the degree to which they are perceived as linguistically-relevant. Participants who were induced into hearing syllables as speech-like did not show any significant differences in performance on our two statistical learning measures compared to participants induced into hearing the syllables as non-linguistic sounds. In addition, participants' ability to linguistically label individual SWS syllables did not predict their statistical learning performance. Taken together, these results provide no strong evidence of a statistical learning advantage for sounds perceived as more speech-like, instead suggesting that statistical learning occurs indiscriminately across auditory stimuli, regardless of their linguistic relevance.

More specifically, on the familiarity rating task, we observed no significant difference in ratings between the speech induced and non-speech induced group, as well as no significant correlation between participants' accuracy in transcribing the SWS syllables and their familiarity rating score. Similarly, on the target detection task, there was no significant difference in the RTs between the induction groups, nor was there a significant correlation between participants' SWS syllable transcription accuracy and the magnitude of their RT prediction effect. Thus, taken together, our results suggest that statistical learning operates largely similarly across physically identical auditory stimuli, regardless of participants' perception of the stimuli as more or less speech-like.

Importantly, we found that the speech induced (SI) group was better at identifying the SWS syllables by their linguistic labels than the non-speech induced (NSI) group, as demonstrated by significantly higher accuracy on the syllable transcription task (30% accuracy for the SI group versus 12% for the NSI). We also found that participants in the SI group rated the syllables as subjectively more speech-like than participants in the NSI group, although the difference in subjective ratings were small. These findings provide a key manipulation check and indicate that our induction task did produce differences in the subjective perception of SWS syllables between the two groups. However, we note that our induction task did not produce a dramatic perceptual transformation of the syllables, as can be found when sentences are used as stimuli (Davis & Johnsrude, 2007; Remez et al., 1981), and was also limited in its generalizability, with no effect on participants' ability to transcribe full sentences. We return to this general point in the Limitations section.

Previous findings in the literature have suggested that statistical learning shows important differences across domains and may be governed by modality- and domain-specific constraints (e.g., Siegelman & Frost, 2015; Siegelman et al., 2017; Frost et al., 2015; Conway et al., 2020; Van Hedger et al., 2022). For example, several findings point to the idea that statistical learning is influenced by the shared resemblance between novel words in the speech stream and existing words in learners' native language, with words that share native language phonotactic patterns being more easily segmented and/or subsequently recognized (Siegelman et al., 2018; Elazar et al., 2022; Finn & Hudson Kam, 2008). Our results provide initial evidence that domain-specific constraints for statistical learning are at least partially attributable to sensory-level processes, and not necessarily to higher-level cognitive mechanisms related to the conceptual categorization of incoming stimuli. For example, networks in auditory cortex may be better equipped to process and encode incoming novel words that have high acoustic overlap with existing words in the learner's lexicon, which in turn could facilitate binding between syllables and lead to observed "linguistic entrenchment" effects (Siegelman et al., 2018). In contrast, the judged linguistic relevance of an ambiguous signal may be a later-occurring, downstream process that does not directly impact statistical learning.

Our approach differed from several previous statistical learning studies in that we did not directly compare learning of speech versus non-speech stimuli (cf. Hoch et al., 2013; Marcus et al., 2007; Ordin et al., 2021; Saffran, 2002; Saffran et al., 1999; Siegelman et al., 2018), which differ in both low-level acoustic features and in communicative relevance. Instead, we assessed the statistical learning of acoustically identical ambiguous stimuli that differed in the degree to which they were subjectively perceived as speech, allowing us to address the more specific question of whether the subjective linguistic value (Berent et al., 2021; Rabagliati et al., 2018) of auditory stimuli—in and of itself—influences statistical learning. To our knowledge, no previous study has directly examined this question in adults. However, there is some relevant prior work in infants, which has examined whether the meaningfulness or communicative relevance of stimuli increases infants' success in learning abstract repetition rules (such as AAB or ABA). Ferguson and Lew-Williams (2016) presented infants with a video prime in which tones were embedded in a natural conversation between two actors, thereby inducing the infants to believe that tones are a communicative signal. In a subsequent rule learning phase, infants who were communicatively primed successfully learned abstract rules from tones, whereas unprimed infants failed to show learning. This finding suggests that infants learn better from stimuli that are communicatively relevant. Supporting this conclusion, a recent meta-analysis of 20 papers (Rabagliati et al., 2018) found that infants are better able to learn abstract repetition rules from stimuli that are communicatively or ecologically meaningful—such as spoken syllables, communicatively primed tones, or natural categories such as dogs or faces—than meaningless stimuli such as geometric shapes or tones. In a follow-up experiment designed to directly test this idea, Rabagliati and colleagues (2018) had infants view either a prime video that portrayed gestures as communicative and meaningful, or a control video, and then exposed them to sequences of gestures following an ABB or ABA pattern. Again, as in Ferguson and Lew-Williams (2016), only infants primed to view gestures as a communicative signal displayed evidence of rule learning. Altogether, these studies suggest that the communicative status of a stimulus enhances abstract rule learning in infants.

In contrast to this general finding in infants, the present results fail to support the idea that the perceived linguistic relevance of auditory stimuli influences or enhances statistical learning in adults. This divergence could potentially be attributed to any number of factors that differ between prior work in infants and the current study, including the population under investigation (adults versus infants), the type of learning (abstract grammatical rule learning versus statistical learning of embedded words in continuous speech), and/or the experimental manipulation used to bias the linguistic relevance of the stimuli. For example, it may be the case that infants show larger differences in learning between communicative and noncommunicative signals compared to adults, in line with the idea that infancy represents a critical period for language acquisition, during which the brain is highly tuned to speech and other communicative signals (Vouloumanos et al., 2010; Vouloumanos & Werker, 2004, 2007; Werker & Hensch, 2015). Another possibility is that findings from abstract grammatical rule learning (e.g., learning of rules such as AAB or ABA) are not directly generalizable to the type of statistical learning under investigation in the current study. Rule learning involves extracting an abstract rule and generalizing to novel instances, whereas statistical learning involves extracting repeating, item-based regularities from unsegmented input, without a generalization component. While these two types of learning appear to be closely related in certain ways (Aslin & Newport, 2012), they may be influenced by different factors and operate under different sets of constraints (Endress & Bonatti, 2007; Endress & Mehler 2009; Peña et al., 2002; Thiessen, 2017).

Finally, we must also consider the possibility that our SWS manipulation did not produce sufficiently diverse percepts of the identical stimuli across individual participants to produce robust differences in statistical learning. Most prior work investigating the processing and intelligibility of SWS have used meaningful sentences (Corcoran et al., 2023; Khoshkhoo et al., 2018; Remez et al., 1981). In contrast, we applied the sine-wave manipulation to isolated syllables, such that participants' perception of the SWS stimuli could not benefit from top-down prediction provided by semantic context. Thus, it is conceivable that even participants who achieved high scores on syllable transcription accuracy may not have experienced a clear speech percept for each syllable. However, a critical point arguing against this possibility is that we did find a significant and highly robust correlation between participants' individual syllable transcription accuracy and overall detection performance for individual syllables in the target detection task. Based on this result, we can conclude that participants experienced real, meaningful variability in their perceptions of the SWS stimuli that was, at minimum, sufficient to robustly predict performance on a separate task. That we did not find similar robust correlations between syllable identification and statistical learning performance suggests that any speech-perception advantage in statistical learning—if it exists at all—is likely to be very small.

The finding that syllable comprehension accuracy predicted overall syllable detection performance in the target detection task is also interesting in and of itself. This result suggests that ability to perceive ambiguous auditory stimuli as more speech-like and the ability to correctly assign linguistic labels to those stimuli facilitate the online identification of the ambiguous stimuli under challenging circumstances, i.e., when the target stimulus is embedded within a continuous stream of similar-sounding sounds. An analogous finding has been reported in the visual domain using a visual search paradigm (Lupyan & Spivye, 2008; Klemfuss et al., 2012). Participants in these studies were presented with arrays of rotated numbers (“2” and “5”), and were asked to indicate for each trial whether the display was homogenous or contained an oddball. Interestingly, participants who were given the linguistic labels or who spontaneously noticed that the shapes were rotated numbers were faster to respond to the arrays compared to participants who were told that the stimuli were abstract shapes. One proposed explanation for this result is that the top-down effects of a linguistic cue may sharpen visual feature detectors, with feedback connections from linguistic representations providing a mechanism for biasing or amplifying activity in perceptual detectors associated with those representations (Lupyan & Spivye, 2008). An alternative explanation is that the benefit of linguistic cues on stimulus identification may occur because language provides a "ready form of efficient coding,” thereby reducing the burden on working memory (Klemfuss et al., 2012). Similar mechanisms operating at both the perceptual and post-perceptual level could also explain the current findings. The ability to perceptually transform a degraded, ambiguous target stimulus into a verbalizable syllable (e.g. “ba”) may have sharpened auditory feature detectors for that sound signal, and may also have facilitated the maintenance of the target stimulus in working memory during the subsequent stream presentation.

4.1. Limitations

As previously alluded to, a limitation in this study was that the speech induction task had only a moderate impact on participants' overall subjective speech perception. As shown in Figure 2A, the speech induction manipulation did not cleanly divide participants into two groups, as some speech-induced participants indicated that they perceived the sounds as relatively un-speechlike, and vice-versa for the non-speech induced participants. In addition, the speech induced group's transcription accuracy of the SWS syllables—while better than the non-speech induced group's—was still fairly low (approximately 30% accuracy). An ideal induction manipulation would have led all the speech-induced participants to accurately perceive the SWS stimuli as speech, and the non-speech induced participants to report hearing the stimuli as non-speech, as was our original intention. This would have allowed for a cleaner comparison between participants speech-induced and non-speech-induced participants, capitalizing on the benefits of an experimental design using random assignment. Because our induction did not result in a clear division between groups, and to account for the continuous, non-binary nature of speech perception, we adopted a complementary approach that tested whether an individual's syllable transcription accuracy predicted their statistical learning performance. However, with this approach there is a possibility that any correlations between transcription performance and statistical learning performance (should they be observed) could be inflated by unintended third variables, such as an individual's general motivation or interest in the experimental tasks. Ultimately, we believe it would be challenging to design a perfectly effective speech induction task when using isolated syllables as SWS stimuli, given their processing cannot benefit from top-down lexical information, which plays an important role in the perceptual learning of distorted speech (Davis et al., 2005). To further probe the role of linguistic relevance in statistical learning, future work could leverage other types of experimental manipulations, such as using priming videos to induce participants into believing that neutral stimuli are a communicative signal (e.g., Ferguson & Lew-Williams, 2016; Rabagliati et al., 2018).

Finally, while the current study demonstrates that overall statistical learning performance is similar as a function of listeners' subjective speech perception, our study design does not allow us determine whether this equivalent performance is supported by a common underlying mechanism or set of mechanisms, or by different mechanisms that depend on speech perception. For example, it is possible that triplets perceived as nonspeech may be segmented and learned as holistic or gestalt-like units, whereas triplets perceived as speech may be learned by extracting sequential syllable patterns—pairs and then triplets—unfolding over time. The theoretical possibility of different mechanisms varying by stimulus material is supported by findings by Siegelman and colleagues (2018), as previously mentioned in the Introduction. This study demonstrated similar overall levels of statistical performance for auditory non-verbal stimuli (everyday sounds) and syllables, which nonetheless belied important differences in the internal consistency of test items between conditions, reflecting different influences on performance that vary by domain. Although we would consider that the possibility of different mechanisms that are equally effective to not necessarily represent the most parsimonious explanation for the current data, the present study design cannot rule it out. Future studies could leverage approaches such as EEG or neuroimaging to examine this possibility directly.

4.2. Conclusions

In summary, our results provide evidence that statistical learning operates largely indiscriminately across auditory stimuli, regardless of the degree to which they are perceived linguistically. In contrast, linguistic perception robustly improves the identification of individual target stimuli embedded in a continuous auditory sequence. These results generally support previous findings of similar statistical learning performance for speech stimuli and non-speech stimuli (Saffran et al., 1999; Saffran, 2002; Siegelman et al., 2018), and raise the possibility that previous demonstrations of the statistical learning advantage for verbal materials (e.g., Hoch et al., 2013; Ordin et al., 2021) may mainly be driven by acoustic differences between the classes of stimuli. These results contribute to the literature on domain-specific versus domain-general contributions to statistical learning, suggesting that statistical learning may be conceptualized as a largely bottom-up mechanism that undiscerningly captures regalities in input regardless of higher-level context.

Data Availability

All data associated with this manuscript are available on Open Science Framework (https://osf.io/jqmxb/?view_only=d7a7d891d2e54a05ad15fe2277dfeb05).

Competing interests

The authors declare no competing interests.

Acknowledgements

This research was supported by a Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (2019-05132) to Laura Batterink.

CRediT Author Statement

Sierra Sweet: writing – original draft, formal analysis, visualization, resources, investigation; **Stephen Van Hedger:** visualization, writing – review and editing, resources, investigation, conceptualization; **Laura Batterink:** conceptualization, writing – review and editing, resources, methodology, supervision, funding acquisition

References

Arciuli, J., & Simpson, I. C. (2011). Statistical learning in typically developing children: The role of age and speed of stimulus presentation. *Developmental Science*, *14*(3), 464–473. https://doi.org/10.1111/j.1467-7687.2009.00937.x

Aslin, R. N. (2017). Statistical learning: A powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews*: *Cognitive Science*, *8*(1–2), 10.1002/wcs.1373. https://doi.org/10.1002/wcs.1373

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science*, *21*(3), 170–176. https://doi.org/10.1177/0963721412436806

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. Cortex, 90, 31–45. https://doi.org/10.1016/j.cortex.2017.02.004

Batterink, L. J., & Paller, K. A. (2019). Statistical learning of speech regularities can occur outside the focus of attention. Cortex, 115, 56–71. https://doi.org/10.1016/j.cortex.2019.01.013

Batterink, L. J., Paller, K. A., & Reber, P. J. (2019). Understanding the Neural Bases of Implicit and Statistical Learning. Topics in Cognitive Science, 11(3), 482–503. https://doi.org/10.1111/tops.12420

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. Journal of Memory and Language, 83, 62–78. https://doi.org/10.1016/j.jml.2015.04.004

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*, 309-312. https://doi.org/10.1038/35002078

Berent, I., de la Cruz-Pavía, I., Brentari, D., & Gervain, J. (2021). Infants differentially extract rules from language. *Scientific Reports*, *11*, Article 20001. https://doi.org/10.1038/s41598-021-99539-8

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S. F., Springer, J. A., Kaufman, J. N., & Possing, E. T. (2000). Human Temporal Lobe Activation by Speech and Nonspeech Sounds. *Cerebral Cortex*, *10*(5), 512–528. https://doi.org/10.1093/cercor/10.5.512

Boersma, Paul & Weenink, David. (2022). *Praat: doing phonetics by computer* (Version 6.2.08). http://www.praat.org/

Boros, M., Magyari, L., Török, D., Bozsik, A., Deme, A., & Andics, A. (2021). Neural processes underlying statistical learning for speech segmentation in dogs. *Current Biology*, *31*(24), 5512-5521.e5. https://doi.org/10.1016/j.cub.2021.10.017

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, *121*(1), 127-132. https://doi.org/10.1016/j.cognition.2011.06.010

Conway, C. M. (2020). How does the brain learn environmental structure? Ten core principles for understanding the neurocognitive mechanisms of statistical learning. *Neuroscience and Biobehavioral Reviews*, *112*, 279–299. https://doi.org/10.1016/j.neubiorev.2020.01.032

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 24-39. https://doi.org/10.1037/0278-7393.31.1.24

Corcoran, A. W., Perera, R., Koroma, M., Kouider, S., Hohwy, J., & Andrillon, T. (2023). Expectations boost the reconstruction of auditory features from electrophysiological responses to noisy speech. *Cerebral Cortex*, *33*(3), 691–708. https://doi.org/10.1093/cercor/bhac094

Covington, N. V., Brown-Schmidt, S., & Duff, M. C. (2018). The Necessity of the Hippocampus for Statistical Learning. *Journal of Cognitive Neuroscience*, *30*(5), 680–697. https://doi.org/10.1162/jocn_a_01228

Darwin, C. (2003). *SWS produced automatically using a script for the PRAAT program*. University of Sussex School of Life Sciences. http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/SWS/

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*, 132–147. https://doi.org/10.1016/j.heares.2007.01.014

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical Information Drives Perceptual Learning of Distorted Speech: Evidence From the Comprehension of Noise-Vocoded Sentences. *Journal of Experimental Psychology: General*, *134*(2), 222–241. https://doi.org/10.1037/0096-3445.134.2.222

Dawson, C., & Gerken, L. (2009). From Domain-Generality to Domain-Sensitivity: 4-Month-Olds Learn an Abstract Repetition Rule in Music That 7-Month-Olds Do Not. *Cognition*, *111*(3), 378–382. https://doi.org/10.1016/j.cognition.2009.02.010

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments. *Journal of Open Source Software*, *8*(85), 5351. https://doi.org/10.21105/joss.05351

Dick, F., Saygin, A. P., Galati, G., Pitzalis, S., Bentrovato, S., D’Amico, S., Wilson, S., Bates, E., & Pizzamiglio, L. (2007). What is Involved and What is Necessary for Complex Linguistic and Nonlinguistic Auditory Processing: Evidence from Functional Magnetic Resonance Imaging and Lesion Data. *Journal of Cognitive Neuroscience*, *19*(5), 799–816. https://doi.org/10.1162/jocn.2007.19.5.799

Elazar, A., Alhama, R. G., Bogaerts, L., Siegelman, N., Baus, C., & Frost, R. (2022). When the “Tabula” is Anything but “Rasa:” What Determines Performance in the Auditory Statistical Learning Task? *Cognitive Science*, *46*(2), e13102. https://doi.org/10.1111/cogs.13102

Emberson, L. L., Conway, C. M., & Christiansen, M. H. (2011). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *Quarterly Journal of Experimental Psychology*, *64*(5), 1021–1040. https://doi.org/10.1080/17470218.2010.538972

Endress, A., & Bonatti, L. (2007). Rapid learning of syllable classes from a perceptually continuous speech stream. *Cognition*, *105*, 247–299. https://doi.org/10.1016/j.cognition.2006.09.010

Endress, A. D., & Mehler, J. (2009). Primitive computations in speech processing. *Quarterly Journal of Experimental Psychology*, *62*(11), 2187–2209. https://doi.org/10.1080/17470210902783646

Ferguson, B., & Lew-Williams, C. (2016). Communicative signals support abstract rule learning by 7-month-old infants*. Scientific Reports, 6*(1), 25434. https://doi.org/10.1038/srep25434

Finn, A. S., & Hudson Kam, C. L. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition*, *108*(2), 477–499. https://doi.org/10.1016/j.cognition.2008.04.002

Fiser, J., & Aslin, R. N. (2001). Unsupervised Statistical Learning of Higher-Order Spatial Structures from Visual Scenes. *Psychological Science*, *12*(6), 499–504. https://doi.org/10.1111/1467-9280.00392

Forest, T. A., Schlichting, M. L., Duncan, K. D., & Finn, A. S. (2023). Changes in statistical learning across development. *Nature Reviews Psychology*, *2*(4), Article 4. https://doi.org/10.1038/s44159-023-00157-0

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *19*(3), 117-125. https://doi.org/10.1016/j.tics.2014.12.010

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128–1153. https://doi.org/10.1037/bul0000210

Gebhart, A. L., Newport, E. L., & Aslin, R. N. (2009). Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychonomic Bulletin & Review*, *16*(3), 486–490. https://doi.org/10.3758/PBR.16.3.486

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53–B64. https://doi.org/10.1016/S0010-0277(00)00132-3

Hoch, L., Tyler, M. D., & Tillmann, B. (2013). Regularity of unit length boosts statistical learning in verbal and nonverbal artificial languages. *Psychonomic Bulletin & Review*, *20*(1), 142–147. https://doi.org/10.3758/s13423-012-0309-8

IEEE. (1969). IEEE Recommended Practice for Speech Quality measurements. *Institute of Electronic Engineers*, New York. https://doi.org/10.1109/IEEESTD.1969.7405210

Khoshkhoo, S., Leonard, M. K., Mesgarani, N., & Chang, E. F. (2018). Neural correlates of sine-wave speech intelligibility in human frontal and temporal cortex. *Brain and Language*, *187*, 83–91. https://doi.org/10.1016/j.bandl.2018.01.007

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–B42. https://doi.org/10.1016/S0010-0277(02)00004-5

Klemfuss, N., Prinzmetal, B., & Ivry, R. (2012). How Does Language Change Perception: A Cautionary Note. Frontiers in Psychology, *3*. https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00078

Liberman, A. M. (1982). On finding that speech is special. *American Psychologist*, *37*(2), 148–167. https://doi.org/10.1037/0003-066X.37.2.148

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Lupyan, G., & Spivey, M. J. (2008). Perceptual processing is facilitated by ascribing meaning to novel stimuli. *Current Biology*, *18*(10), R410–R412. https://doi.org/10.1016/j.cub.2008.02.073

Marcus, G. F., Fernandes, K. J., & Johnson, S. P. (2007). Infant Rule Learning Facilitated by Speech. *Psychological Science, 17*(5), 387-391. https://doi.org/10.1111/j.1467-9280.2007.01910.x

Marcus, G. F., & Rabagliati, H. (2008). In J. Colombo, P. McCardle & L. Freund (Eds.), *Infant Pathways to Language: Methods, Models and Research Directions.* Lawrence Erlbaum Associates.

Mitchel, A. D., Christiansen, M. H., & Weiss, D. J. (2014). Multimodal integration in statistical learning: Evidence from the McGurk illusion. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00407

Marno, H., Farroni, T., Vidal Dos Santos, Y., Ekramnia, M., Nespor, M., & Mehler, J. (2015). Can you see what I am talking about? Human speech triggers referential expectation in four-month-old infants. *Scientific Reports*, *5*. https://doi.org/10.1038/srep13594

Moore, D. R. (2000). Auditory neuroscience: Is speech special? *Current Biology*, *10*(10), R362–R364. https://doi.org/10.1016/S0960-9822(00)00479-6

Moreau, C. N., Joanisse, M. F., Mulgrew, J., & Batterink, L. J. (2022). No statistical learning advantage in children over adults: Evidence from behaviour and neural entrainment. Developmental Cognitive Neuroscience, 57, 101154. https://doi.org/10.1016/j.dcn.2022.101154

Narain, C., Scott, S. K., Wise, R. J. S., Rosen, S., Leff, A., Iversen, S. D., & Matthews, P. M. (2003). Defining a Left-lateralized Response Specific to Intelligible Speech Using fMRI. *Cerebral Cortex*, *13*(12), 1362–1368. https://doi.org/10.1093/cercor/bhg083

Ogg, M., & Slevc, L. R. (2019). Acoustic Correlates of Auditory Object and Event Perception: Speakers, Musical Timbres, and Environmental Sounds. *Frontiers in Psychology, 10*. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01594

Ordin, M., Polyanskaya, L., & Samuel, A. (2021). An evolutionary account of intermodality differences in statistical learning. *Annals of New York Academy of Science*, *1486*(1), 76–89. https://doi.org/10.1111/nyas.14502

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Parviainen, T., Helenius, P., & Salmelin, R. (2005). Cortical differentiation of speech and nonspeech sounds at 100 ms: Implications for dyslexia. *Cerebral Cortex*, *15*(7), 1054–1063. https://doi.org/10.1093/cercor/bhh206

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-Driven Computations in Speech Processing. *Science*, *298*(5593), 604–607. https://doi.org/10.1126/science.1072901

Rabagliati, H., Ferguson, B., & Lew-Williams, C. (2018). The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental Science, 22*(1). https://doi.org/10.1111/desc.12704

Rabagliati, H., Senghas, A., Johnson, S., & Marcus, G. F. (2012). Infant Rule Learning: Advantage Language, or Advantage Speech? *PLOS ONE*, *7*(7), e40517. https://doi.org/10.1371/journal.pone.0040517

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). https://doi.org/10.1111/desc.12593

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech Perception without Traditional Speech Cues. *Science, 212*(4497), 947–950. https://doi.org/10.1126/science.7233191

Saffran, J. R. (2002). Constraints on Statistical Language Learning. *Journal of Memory and Language, 47*(1), 172-196. https://doi.org/10.1006/jmla.2001.2839

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*, *35*(4), 606–621. https://doi.org/10.1006/jmla.1996.0032

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*(2), 101–105. https://doi.org/10.1111/j.1467-9280.1997.tb00690.x

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52. https://doi.org/10.1016/S0010-0277(98)00075-4

Saffran, J. R., Pollak, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, *105*(3), 669–680. https://doi.org/10.1016/j.cognition.2006.11.004

Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., & Turk-Browne, N. B. (2014). The Necessity of the Medial Temporal Lobe for Statistical Learning. *Journal of Cognitive Neuroscience*, *26*(8), 1736–1747. https://doi.org/10.1162/jocn_a_00578

Schlichting, M. L., Guarino, K. F., Schapiro, A. C., Turk-Browne, N. B., & Preston, A. R. (2017). Hippocampal Structure Predicts Statistical Learning and Associative Inference Abilities during Development. *Journal of Cognitive Neuroscience*, *29*(1), 37–51. https://doi.org/10.1162/jocn_a_01028

Schmalz, X., Biurrun Manresa, J., & Zhang, L. (2023). What is a Bayes factor? *Psychological Methods*, *28*(3), 705-718. https://doi.org/10.1037/met0000421

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*(12), 2400–2406. https://doi.org/10.1093/brain/123.12.2400

Seitz, A. R., Kim, R., Van Wassenhove, V., & Shams, L. (2007). Simultaneous and Independent Acquisition of Multisensory and Unisensory Associations. *Perception*, *36*(10), 1445–1453. https://doi.org/10.1068/p5843

Shufaniya, A., & Arnon, I. (2018). Statistical Learning Is Not Age-Invariant During Childhood: Performance Improves With Age Across Modality. *Cognitive Science*, *42*(8), 3100–3115. https://doi.org/10.1111/cogs.12692

Shultz, S., & Vouloumanos, A. (2010). Three-Month-Olds Prefer Speech to Other Naturally Occurring Signals. *Language Learning and Development*, *6*(4), 241–257. https://doi.org/10.1080/15475440903507830

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711). https://doi.org/10.1098/rstb.2016.0059

Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition, 177*, 198-213. https://doi.org/10.1016/j.cognition.2018.04.011

Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, *81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001

Silva, D. M. R., & Bellini-Leite, S. C. (2020). Cross-modal correspondences in sine wave: Speech versus non-speech modes. *Attention, Perception, & Psychophysics, 82*(3), 944–953. https://doi.org/10.3758/s13414-019-01835-z

Thiessen, E. D., (2012). Effects of Inter- and Intra-modal Redundancy on Infants' Rule Learning. *Language Learning and Development, 8*(3), 197-214. https://doi.org/10.1080/15475441.2011.583610

Thiessen, E. D. (2017). What's statistical about learning? Insights from modelling statistical learning as a set of memory processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160056. https://doi.org/10.1098/rstb.2016.0056

Van Hedger, S. C., Johnsrude, I. S., & Batterink, L. J. (2022). Musical instrument familiarity affects statistical learning of tone sequences. Cognition, 218, 104949. https://doi.org/10.1016/j.cognition.2021.104949

Vouloumanos, A., Hauser, M. D., Werker, J. F., & Martin, A. (2010). The Tuning of Human Neonates' Preference for Speech. *Child Development*, 81(2), 517–527. https://doi.org/10.1111/j.1467-8624.2009.01412.x

Vouloumanos, A., Kiehl, K. A., Werker, J. F., & Liddle, P. F. (2001). Detection of Sounds in the Auditory Stream: Event-Related fMRI Evidence for Differential Activation to

Speech and Nonspeech. *Journal of Cognitive Neuroscience*, *13*(7), 994–1005. https://doi.org/10.1162/089892901753165890

Vouloumanos, A., & Werker, J. F. (2004). Tuned to the signal: The privileged status of speech for young infants. *Developmental Science*, *7*(3), 270-276. https://doi.org/10.1111/j.1467-7687.2004.00345.x

Vouloumanos, A., & Werker, J. F. (2007). Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental Science*, *10*(2), 159–164. https://doi.org/10.1111/j.1467-7687.2007.00549.x

Werker, J. F., & Hensch, T. K. (2015). Critical Periods in Speech Perception: New Directions. *Annual Review of Psychology*, *66*(1), 173–196. https://doi.org/10.1146/annurev-psych-010814-015104