CORTEX 115 (2019) 56–71

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Research Report

Statistical learning of speech regularities can occur outside the focus of attention

Laura J. Batterink [a,b,*] *and Ken A. Paller* [b]

[a] *Western University, Department of Psychology, Brain & Mind Institute, London, ON, Canada*
[b] *Northwestern University, Department of Psychology, Evanston, IL, USA*

ARTICLE INFO

Article history:
Received 3 April 2018
Reviewed 14 June 2018
Revised 6 August 2018
Accepted 10 January 2019
Action editor Sonja Kotz
Published online 28 January 2019

Keywords:
Statistical learning
Speech segmentation
Attention
Memory
Neural entrainment

ABSTRACT

Statistical learning, the process of extracting regularities from the environment, plays an essential role in many aspects of cognition, including speech segmentation and language acquisition. A key component of statistical learning in a linguistic context is the perceptual binding of adjacent individual units (e.g., syllables) into integrated composites (e.g., multisyllabic words). A second, conceptually dissociable component of statistical learning is the memory storage of these integrated representations. Here we examine whether these two dissociable components of statistical learning are differentially impacted by top-down, voluntary attentional resources. Learners' attention was either focused towards or diverted from a speech stream made up of repeating nonsense words. Building on our previous findings, we quantified the online perceptual binding of individual syllables into component words using an EEG-based neural entrainment measure. Following exposure, statistical learning was assessed using offline tests, sensitive to both perceptual binding and memory storage. Neural measures verified that our manipulation of selective attention successfully reduced limited-capacity resources to the speech stream. Diverting attention away from the speech stream did not alter neural entrainment to the component words or post-exposure familiarity ratings, but did impact performance on an indirect reaction-time based memory test. We conclude that theoretically dissociable components of statistically learning are differentially impacted by attention and top-down processing resources. A reduction in attention to the speech stream may impede memory storage of the component words. In contrast, the moment-by-moment perceptual binding of speech regularities can occur even while learners' attention is focused on a demanding concurrent task, and we found no evidence that selective attention modulates this process. These results suggest that learners can acquire basic statistical properties of language without directly focusing on the speech input, potentially opening up previously overlooked opportunities for language learning, particularly in adult learners.

* *Corresponding author*. Western University, Department of Psychology, Brain & Mind Institute, London, ON, N6A 5B7, Canada.
E-mail addresses: lbatter@uwo.ca (L.J. Batterink), kap@northwestern.edu (K.A. Paller).

https://doi.org/10.1016/j.cortex.2019.01.013
0010-9452/

1. Introduction

Listening to an unfamiliar language can be a disorienting experience. Natural speech is a continuous stream of sound, with no reliable pauses between individual words (Lehiste, 1960). One of the first steps in acquiring an unfamiliar language is the discovery of word boundaries in this continuous speech stream. This challenge is thought to be at least partially solved through *statistical learning*, which refers to acquisition of statistical structure in the environment. In spoken language, adjacent syllables within words co-occur more often than syllables that are adjacent but cross word boundaries. Gradually gaining knowledge of these syllable co-occurrences is one way for learners to discover word boundaries in continuous speech, as has been demonstrated convincingly by studies using novel artificial speech streams (e.g., Saffran, Aslin & Newport, 1996; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997; Saffran, Newport & Aslin, 1996).

As we have previously proposed (Batterink & Paller, 2017a, b), statistical learning performance can be conceptualized as comprising at least two dissociable components: (1) perceptual binding and (2) subsequent memory storage and retrieval. Perceptual binding involves a transition from the perception and encoding of raw individual stimulus units to that of larger integrated items. For example, in the case of word segmentation, learners exposed to an unfamiliar language initially perceive a sequence of individual syllables, rather than the organized sequence of coherent words that fluent speakers perceive. With sufficient exposure to a language, learners' initial perception of these smaller syllable units is gradually transformed to that of larger word units. The perceptual process of identifying words in continuous speech occurs *online*, during exposure to input, and may be considered the central challenge of statistical learning. A second key component of statistical learning involves the storage and subsequent retrieval of these extracted representation in long-term memory. These memory-related processes can in one sense be considered peripheral to the central task of statistical learning. Nonetheless, both the perceptual binding component and memory component critically influence all typical measures of statistical learning performance.

Previous studies of statistical learning have generally been unable to distinguish between these two conceptually distinct components of learning. This limitation is largely due to the experimental approach that is most commonly used to investigate statistical learning, in which learning is assessed offline following an exposure period to visual or auditory regularities. For example, a typical auditory statistical learning paradigm involves exposing participants to a continuous speech stream of repeating three-syllable nonsense "words." This exposure period is then followed by a subsequent recognition test in which participants discriminate between words and novel foils through a two-alternative forced choice measure (e.g., Saffran et al., 1997; Saffran, Newport, et al., 1996). Reaction-time-based tests have also been used, in which participants are asked to identify a target stimulus presented in a continuous stream (Batterink, Reber, Neville, et al., 2015b; Batterink, Reber, & Paller, 2015a; Franco, Eberlen, Destrebecqz, Cleeremans, & Bertels, 2015; Kim, Seitz, Feenstra, & Shams, 2009; Turk- Browne et al., 2005). Faster response times to predictable targets (i.e., those that occurs in positions 2 or 3 of a repeating triplet) compared to unpredictable targets (i.e., those that occur in position 1) provide evidence of statistical learning. Reaction-time tests may provide a more sensitive measure of learning that can potentially capture knowledge above and beyond what is reflected by recognition measures (Batterink et al., 2015a, b). Nonetheless, both types of test are administered after the exposure period and assess only the final outcome of learning. Thus, these offline tasks cannot disentangle contributions from online perceptual binding versus subsequent memory storage and retrieval processes to statistical learning.

In contrast to these offline behavioral tests, neural measures can assess statistical learning online, *during* exposure to structured input. In particular, several event-related potential (ERP) components appear to be modulated as a function of statistical learning during exposure. For example, Cunillera, Toro, Sebastián-Gallés, and Rodríguez-Fornells (2006, Cunillera, Càmara et al., 2009) compared ERPs to words and random syllables in continuous speech. Words elicited a larger N400 relative to nonwords, and interestingly this effect emerged very quickly, after only 1 min of exposure (Cunillera et al., 2009). Similarly, De Diego Balaguer, Toro, Rodriguez-Fornells, and Bachoud-Lévi (2007) found an increase in N400 amplitudes to words in continuous speech as a function of exposure, with larger N400 amplitudes observed in the second minute of exposure compared to the first minute. This N400 effect was also associated with better word recognition on a post-learning task. Sanders, Newport, and Neville (2002) pretrained learners on six nonsense words prior to exposing them to a continuous stream of the concatenated words. In participants showing the greatest behavioral evidence of word learning, word onsets elicited a larger N100 after compared to before training, suggesting that the N100 may index segmentation or recognition of previously learned words in continuous speech. Finally, Abla and colleagues found that both the N100 and N400 components tracked statistical learning of repeating tone triplets, or "tritone words," with tones in the initial position eliciting larger N100 and N400 effects compared to second and third position tones. In sum, ERP studies have shown that statistical learning is associated with the emergence of specific electrophysiological markers. The N400 appears to be the most reliable index of statistical learning across studies, and may represent the construction of a pre-lexical trace for new words (De Diego Balaguer et al., 2007). ERP effects appear to emerge rapidly during exposure to structured input and provide insight into both the underlying mechanisms and the time course of statistical learning.

In addition to ERPs, a "frequency-tagging" approach has also been used to track statistical learning online (Buiatti, Pena, & Dehaenelambertz, 2009; Kabdebon, Pena, Buiatti, & Dehaene-Lambertz, 2015; Batterink & Paller, 2017a, b). This approach takes advantage of the neural steady-state response, the tendency of the brain to entrain or oscillate at the same frequency as an ongoing rhythmic stimulus. In addition to reflecting bottom-up sensory processing, neural entrainment is sensitive to internally driven stimulus integration processes, reflecting abstract, higher-level features such as syntactic rules (Ding et al., 2016) or imagined metric

beats (Nozaradan et al., 2011). Ding and colleagues presented isochronous monosyllabic words in both a natural and artificial language and observed neural entrainment not only at the syllabic rate, but also at the level of phrases and sentences, reflecting learners' knowledge of hierarchical syntactic rules. Similarly, when participants were instructed to imagine a binary or ternary metric beat while listening to a sequence of isochronous tones, increased neural entrainment was observed at the corresponding imagined beat frequencies (Nozaradan et al., 2011).

Because of its sensitivity to perpetual integration processes, the frequency tagging method offers a particularly promising means to quantify the perceptual binding component of statistical learning. The central assumption here is that as learners gradually discover words in continuous speech, their brains should progressively entrain more strongly at the word frequency compared to the raw syllable frequency, representing a shift in perception from syllables to words. This idea appears to be supported by several recent studies. Buiatti et al. (2009) presented participants with a speech stream composed of trisyllabic artificial words constructed according to a nonadjacent *AXC* rule structure, in which the third syllable of a word was predicted by the first syllable. Words were either separated by a 25-msec subliminal pause or concatenated together continuously. Interestingly, a peak in neural entrainment at the word frequency was observed only when the *AXC* words were separated by the subliminal pauses, suggesting that the pauses facilitated learning of the nonadjacent dependencies and corresponding word extraction. This condition was also accompanied by a suppression in entrainment at the syllable frequency, reflecting that syllables were no longer processed in isolation but rather linked together with neighboring syllables to form coherent words. Kabdebon et al. (2015) exposed infants to a similar artificial stream of *AXC* words separated by 25-msec subliminal pauses, and found that infants also showed neural entrainment at the word frequency.

Building on these findings, in a recent study we used neural entrainment to track "pure" statistical learning, presenting concatenated words without any subliminal pauses that may facilitate segmentation. Even without the insertion of pauses, we found that neural entrainment tracks the extent to which learners bind neighboring syllables into the underlying component words of the speech stream (Batterink & Paller, 2017a, b). Specifically, we computed EEG-based neural entrainment at the frequency of the repeating words relative to that of individual syllables. The ratio of neural entrainment to words versus syllables (which we termed the "word learning index," or *WLI*) distinguishes between structured and random input, tracks the progression of learning over time, and predicts performance on a subsequent offline reaction-time test of statistical learning. Thus, this neural entrainment measure appears to reflect learners' perceptual sensitivity to the hidden component words within the structured speech stream, without requiring subsequent behavioral testing. Applied to future studies, this measure may be a valuable tool to address questions about statistical learning, such as the necessary versus sufficient conditions under which this type of learning can proceed.

In the present study, we addressed one such outstanding question: the role of top-down, voluntary attention to the perceptual binding component and subsequent memory processes underlying statistical learning. Only a small subset of all possible stimuli in the environment can be selected (*attention*) and maintained in an active state over time (*working memory*; cf. Fougnie et al., 2008). Attention and working memory are limited in capacity, such that allocation of these resources towards some information comes at the cost of processing other information. In a typical listening environment, learners may choose to focus their attention and limited processing resources on a speech stream, or on other competing stimuli and tasks. Thus, it is important to determine whether statistical learning occurs automatically or is impacted by the voluntary allocation of these resources. We sought to address this issue by manipulating the focus of learners' voluntary attention using a dual task manipulation. We aimed to answer several interrelated questions. First, can learners become sensitive to the component words in structured speech even in the absence of focused attention to the speech stream? Relatedly, does focused attention to speech facilitate learners' ability to bind neighboring syllables into accurate component words in continuous speech? Finally, does focused attention to the speech stream have different effects on online perceptual word binding and subsequent memory storage processes? Resolving these questions would lead to a better understanding of the mechanisms of statistical learning. These results would also provide important insight into the external conditions under which statistical learning does or does not occur, yielding information that may have valuable practical implications for language acquisition.

Although no prior research has looked at these questions specifically, a small number of previous studies have investigated the role of selective attention in overall levels of statistical learning, as assessed by performance on offline tests. Several of these studies suggest that statistical learning may require some degree of selective attention to the stimuli. Using a typical auditory statistical learning paradigm, Toro, Sinnet, and Soto-Faraco (2005) found a dramatic decrease in offline recognition of words versus nonwords when participants' attention was diverted to an unrelated, concurrent task, relative to conditions of passive listening. Under some types of attentional manipulations, performance declined to chance levels. Using a similar design, Palmer and Mattys (2016) demonstrated an impairment in performance on a forced-choice recognition task under conditions of high attentional load. Attention has also been observed to impact statistical learning in the visual domain (Turk-Browne, Junge, & Scholl, 2005). In this study, learners' attention was directed to one of two interleaved stimulus sets of repeating shape triplets. Whereas robust statistical learning was observed to the attended stream, performance was at chance level for the unattended stream across different tasks. Taken together, these results suggest that statistical learning under conditions of reduced attention is seriously compromised and may often fail to occur at all; in order for statistical learning to occur, some level of attentional resources must be directed to the input stream.

In contrast, in a more recent visual statistical learning study with a design comparable to that used by Turk-Browne et al. (2005), equivalent learning was observed for both the attended and unattended stimulus set (Musz, Weber, & Thompson-Schill, 2015). Not even a weak effect of attention was present across several different experimental manipulations, suggesting that statistical learning can occur even when attention is actively directed away from unattended input and towards competing stimuli. At present, the reason for the discrepancy in results between these two studies is not obvious, given the high similarity in experimental procedures. In the auditory domain, statistical learning has also been shown to occur to a speech stream played in the background, while participants were actively engaged in a picture-drawing task (Saffran et al., 1997). Although this study was not specifically designed to manipulate attention, these results suggest that statistical learning of speech regularities can occur to input that is processed outside of learners' focus of attention. Finally, Fernandes, Kolinsky, and Ventura (2010) found that high attentional load had no impact on statistical learning of words with high transitional probabilities, but negatively influenced learning of words with less salient statistical cues.

In sum, there are conflicting results regarding the role of attention in statistical learning as assessed through offline tasks, and the question of whether focused attention to input is necessary for such learning has not yet been satisfactorily resolved. Further, these prior studies have been unable to address the additional question of whether attention and top-down resources may play different roles in different components of statistical learning—namely perceptual binding and subsequent memory-related processes—because of their reliance on offline measures.

To investigate the role of focused attention and associated central, limited-capacity resources on these different components of statistical learning, we used our previously described neural entrainment measure to quantify the perceptual binding of component words during exposure to structured speech (Batterink & Paller, 2017a, b). To assess overall levels of statistical leaning performance, we administered both a direct, familiarity-based rating measure and an indirect, reaction-time based measure after the exposure period (Batterink, Reber, & Paller, 2015a; 2015b, Batterink & Paller, 2017a, b). Our familiarity-rating task required participants to provide ratings of words, part-words, and non-words, providing a measure of explicit knowledge of the underlying component words of the stream. Our reaction-time based measure involved a target detection task, in which participants were required to make speeded responses to syllable targets in short streams of continuous speech (e.g., Batterink, Reber, Neville, & Paller, 2015b, 2015a; Franco et al., 2015; Kim et al., 2009; Turk-Browne et al., 2005). Faster reaction times (RTs) index more efficient processing of predictable syllable targets (i.e., those occurring in the later positions of a trisyllabic word). In previous work, we found that our online neural entrainment measure predicted performance on the target detection task, with greater relative entrainment at the word frequency correlating with a larger reaction time (RT) effect on the target detection task (Batterink & Paller, 2017a, b).

Using a standard statistical learning speech segmentation task, in which learners are presented with a continuous speech stream of repeating three-syllable nonsense words, we compared neural entrainment and behavioral performance under two processing conditions. Participants assigned to a "Full Attention" condition were instructed to focus fully on the speech stream, whereas other participants in a "Divided Attention" condition were required to perform a demanding concurrent visual task while the speech stream was presented, such that limited-capacity resources to the speech input would be reduced. Based on prior results (Toro et al., 2005; Palmer & Mattys, 2016), we hypothesized that diverting the focus of learners' attention away from the speech stream would decrease performance on the explicit familiarity-rating task. Given lack of prior evidence, we remained agnostic about whether our attentional manipulation would modulate neural entrainment to the underlying words and performance on the target detection task. We also examined the time course of neural entrainment across the exposure period to test whether reduced attention to the speech stream delays the progression of learning, even if it does not prevent it entirely. This possibility has been proposed to occur (Fernandes et al., 2010; Toro et al., 2005), but has not been directly tested. Any of these different patterns of results would inform our understanding of the role of attention and limited-capacity resources in the perceptual binding of speech regularities and memory-related components of statistical learning.

2. Material and methods

2.1. Participants

A total of 49 participants were run (mean age = 21.8 years, SD = 2.8 years, 37 female/12 male), assigned randomly to the Divided Attention condition (n = 25) or the Full Attention condition (n = 21). Three additional participants were excluded from the Full Attention condition, due to either technical failures with EEG recording (n = 2) or extremely poor target-detection performance (<10% targets detected; n = 1).

All participants were fluent English speakers and had no history of neurological problems. Experiments were undertaken with the understanding and written consent of each participant. Participants were compensated at $10/h for their time.

2.2. Stimuli

Syllables contributing to the speech stream were recorded by a male native English speaker using neutral intonation and no co-articulation between syllables. Individual sound files, each containing a single syllable, were spliced from the recordings. The beginning of each sound file coincided with the precise onset of the syllable. Syllables had an approximate duration of 220–250 msec from onset to offset and were equated for perceived volume. The continuous speech stream was created by concatenating the individual syllables together in a preset order, as described more fully below, with a constant stimulus onset asynchrony between each syllable.

For the visual 3-back task, a total of 10 unique kaleidoscope images were used (see Fig. 1 for example images). These images were selected because they are difficult to label verbally

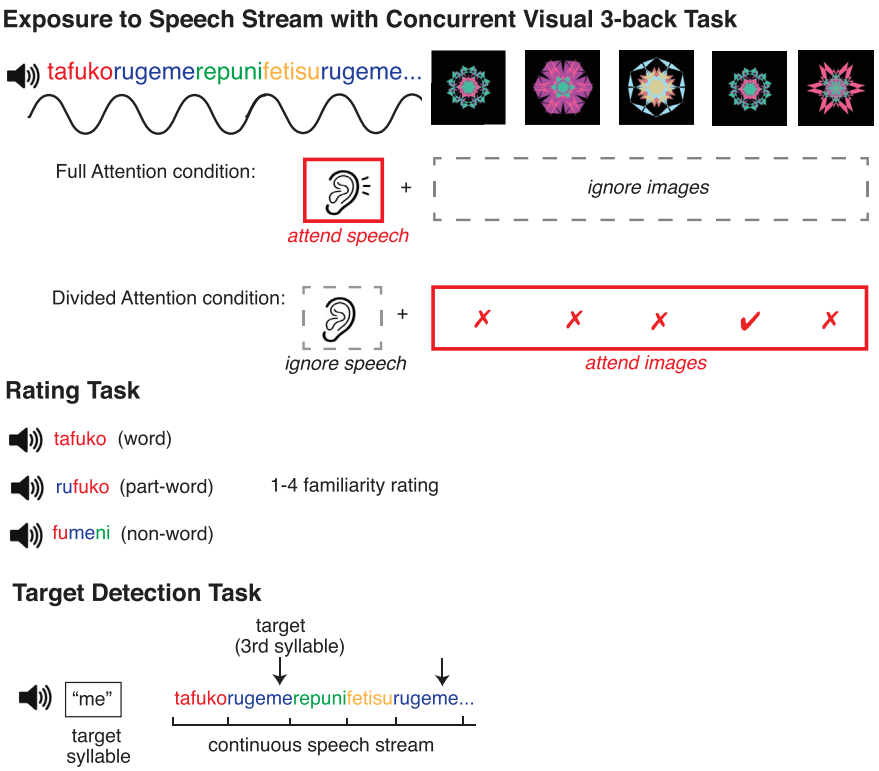

Fig. 1 – Summary of experimental design. The exposure task in the structured condition consisted of 12 min of continuous auditory exposure to four repeating nonsense words. Participants in the Full Attention condition were instructed to attend the speech stream while passively viewing the kaleidoscope images, whereas participants in the Divided Attention condition were instructed to perform a demanding 3-back task on the visual images, while ignoring the speech stream. Our online neural entrainment measure was used to assess statistical learning during the exposure task. After exposure, statistical learning was assessed using the familiarity-rating and target detection tasks.

and thus less likely to tax phonetic or linguistic attentional resources (Voss & Paller, 2009), allowing us to examine the effects of domain-general attention and limited-capacity resources on auditory statistical learning. Images were created by overlaying three opaque hexagons of different colors and performing three rounds of side bisection and random deflection on each. All images were overlaid over a black background on a computer monitor placed approximately 120 cm in front of the participant, presented within the upper half of the display. For participants in the Divided Attention condition, thumbs-up and thumbs-down images were also presented in the lower half of the display as a method of feedback on performance.

2.3. *Procedure*

Fig. 1 shows a visual summary of the experimental design. Auditory stimuli were presented at a comfortable listening level (approximately 70–75 db) from two speakers placed approximately 120 cm in front of the participant.

2.3.1. *Exposure task*

During the exposure task, participants were exposed to an auditory speech stream made up of repeating nonsense words. Participants in both attention conditions were informed that they would hear a "nonsense language," but they were not given any information about the structure of the language, such as the number of component words or the number of syllables per word. Syllables were grouped into 4 repeating trisyllabic words (e.g., *tafuko*, *rugeme*, *repuni*, and *fetisu*), with the transitional probability between syllables higher within words (1.0) than between words (.33). For example, a transitional probability of 1.0 for a word, like *tafuko*, means that every *ta* in the stream was followed by *fu* and every *fu* was followed by *ko*; in contrast, *ko* was equally likely to be followed by *ru*, *re* or *fe* (words were not allowed to repeat). A total of 2400 syllables (corresponding to 800 "words") were presented at a constant rate for each participant (range across participants: 260 msec/syllable – 300 msec/syllable). We originally intended to use the same stimulus presentation rate as our earlier study for all participants (300 msec/syllable; Batterink & Paller, 2017a, b), but a computer hardware upgrade made partway through data collection unexpectedly produced faster stimulus presentation rates for approximately half of the participants. This small change in stimulus presentation rate affected equal numbers of participants in both groups (divided attention group: faster rate $n = 12$, slower rate $n = 13$; full attention group: faster rate $n = 10$; slower rate $n = 11$), and we confirmed statistically that stimulus presentation rate did not significantly differ between the two groups [$t(44) = .033$, $p = .97$]. The syllable stream was broken up into 3 blocks. Each block contained exactly 800

syllables, with a duration of approximately 4 min; blocks 2 and 3 did not begin with a word-initial syllable and thus block onset did not provide any additional segmentation cues. Participants were given a brief break after each block and allowed to resume the task whenever ready.

To ensure that idiosyncratic perceptual differences between syllables could not drive group-level statistical learning performance, the assignment of individual syllables to the first, second and third positions of each word was counterbalanced across participants, resulting in three different counterbalancing condition. For example, counterbalance order 1 contained the word *tafuko*, counterbalance order 2 the word *fukota*, and counterbalance order 3 the word *kotafu*.

2.3.2. *Concurrent visual 3-back task*

While listening to the speech stream, all participants concurrently viewed a sequence of kaleidoscope images presented on a computer monitor. The same image was never repeated consecutively. The duration of each image was randomized at both the individual and trial levels, with each image presented for a random time interval from 2.4 to 5 sec before being replaced by the subsequent image. This randomization procedure ensured that the onset of visual stimuli could not provide a segmentation cue and would not systematically influence the auditory EEG response at the group level. Image order was predetermined and consistent across all participants. Approximately 25% of the images represented a 3-back match (i.e., the same image presented 3 previously), with the remaining 75% of trials representing 3-back non-match.

Participants in the Full Attention condition were instructed to attend to the auditory speech stream and to simply view the images passively. In contrast, participants in the Divided Attention condition were required to perform an attention-demanding 3-back task on the sequence of images, which involved indicating whether the current stimulus matched the image presented 3 steps earlier in the sequence (Fig. 1). They were advised that there would be a nonsense language playing in the background, but that they should ignore the language and concentrate on the visual task. They received feedback on their performance after every trial. If they responded correctly within 2 sec of the image onset, a thumbs-up icon appeared below the kaleidoscope image. If they responded incorrectly, or failed to respond within 2 sec of the image onset, a thumbs-down icon appeared below the kaleidoscope. Participants in both conditions were instructed to keep their eyes focused on the center of the computer monitor at the kaleidoscope images and to minimize large eye movements.

2.3.3. *Familiarity-rating task*

Following exposure to the structured stream, participants completed a familiarity-rating task designed to assess explicit memory of the nonsense words, as we have employed previously (Batterink & Paller, 2017a, b).

On each trial, participants were presented with one of three types of auditory stimuli: a word from the language that had been previously presented 800 times during the Exposure task (e.g., *tafuko*), a part-word that consisted of a syllable pair from a word from the language plus an additional syllable (e.g., *rufuko*), or a non-word that consisted of three syllables from the language that were never paired together within a word (e.g., *fumeni*). Without time pressure, participants were asked to rate on a 1–4 scale how familiar the stimulus sounded based on the language that they had just heard, with 1 indicating "very unfamiliar" and 4 indicating "very familiar." A total of 12 trials were presented, consisting of 4 words, 4 part-words, and 4 non-words.

2.3.4. *Target detection task*

Finally, participants completed a speeded target detection task (Batterink et al., 2015a; 2015b; Batterink & Paller, 2017a, b). On each trial, participants were presented with a speech stream containing the four words from the structured language presented four times each in pseudorandom order, which was shorter but otherwise similar to the speech stream presented during the Exposure task. For each stream, participants were required to detect a specific target syllable. Both RT and accuracy were emphasized. Each of the 12 syllables of the structured syllable inventory served as the target syllable three times, for a total of 36 streams. The order of the 36 streams was randomized for each participant. Each stream contained 4 target syllables, providing a total of 48 trials in each of the three-syllable conditions (word-initial, word-medial, and word-final). At the beginning of each trial, participants pressed "Enter" to listen to a sample of the target syllable. The stimulus stream was then initiated. Stimulus timing parameters were identical to those in the Exposure task. Based on our previous findings (Batterink et al., 2015a; 2015b, Batterink & Paller, 2017a, b), we expect graded RT effects as a function of syllable position. Syllable targets that occur in the final position of a word should elicit faster RTs, indexing facilitation due to statistical learning.

2.4. *Behavioral data analysis*

For participants in the Divided Attention condition, performance on the 3-back task was quantified using *d'*, with hit rate quantified as the proportion of 3-back matches that were classified correctly and false alarm rate quantified as the proportion of non-matches that were incorrectly classified as hits. This measure was used rather than overall accuracy as there were more non-match trials than match trials, and thus simply responding "non-match" to all trials would lead to a performance of 75% accuracy.

On the familiarity-rating task, ratings were analyzed using a repeated-measures ANOVA with word category (word, part-word, non-word) as a within-participants factor and attention condition (full, divided) as a between-participants factor. For correlational analyses, performance was also quantified by subtracting the average rating to foil items (both part-words and non-words) from the average rating to words, for each participant. Perfect sensitivity on this "familiarity rating score" would be a score of 3, with values above 0 providing evidence of learning. As an additional measure of performance on this task, RTs were analyzed using a second repeated-measures ANOVA with the same factors as above. Median RTs were computed within each word category and participant to reduce the influence of outliers.

For the target-detection task, responses that occurred within 1200 msec after a target were considered to be hits, whereas responses that occurred anytime other than within 0–1200 msec of a target were considered to be false alarms; this is the same criterion used in all our past studies (Batterink et al., 2015a, b, Batterink & Paller, 2017a, b). Mean RTs to detected targets (hits) were calculated for each syllable condition (word-initial, word-medial, and word-final) for each participant; mean rather than median was used as a measure of central tendency in this analysis given that RTs longer than 1200 msec were already excluded according to our "hit" criterion. RTs were analyzed using a repeated-measures ANOVA with syllable position (initial, medial, final) as a within-participants factor, and attention condition (full, divided) as a between-participants factor. Planned contrasts were used to examine whether RTs decreased linearly as a function of syllable position. Performance was further quantified through an "RT prediction effect," computed as the proportion of RT decrease to third position targets relative to initial position targets $[(RT_1 - RT_3)/RT_1]$. Because decreases in RTs are not independent of the overall speed of response (cf. Siegelman, Bogaerts, Kronenfeld, & Frost, 2017), this computation adjusts for potential differences in baseline RTs between individuals, allowing us to compare statistical learning across individuals with different RT baselines. Larger positive values on the RT prediction effect indicate greater facilitation.

2.5. EEG recording and analysis

During both the Exposure Task and the Target Detection task, EEG was recorded with a sampling rate of 512 Hz from 64 Ag/AgCl-tipped electrodes attached to an electrode cap using the 10/20 system. Recordings were made with the Active-Two system (Biosemi, Amsterdam, The Netherlands). Additional electrodes were placed on the left and right mastoid, at the outer canthi of both eyes, and below both eyes. Scalp signals were recorded relative to the Common Mode Sense (CMS) active electrode and then re-referenced off-line to the algebraic average of the left and right mastoid.

2.5.1. Quantification of neural entrainment

EEG neural entrainment analyses were carried out using EEGLAB (Delorme & Makeig, 2004) and followed the same general procedure as in our previous study (Batterink & Paller, 2017a, b).

EEG data acquired during the Exposure task were band-pass filtered from .1 to 30 Hz. Data from each block were timelocked to the onset of each word and extracted into epochs of 10.8 sec, corresponding to the duration of 12 trisyllabic words or 36 syllables (with no pre-stimulus interval). This procedure yielded epochs overlapping for 11/12 of their length. We employed an automatic artifact rejection procedure designed to remove only data containing large artifacts, based on threshold amplitude values adjusted individually for each participant (range = 200–300 μV). Data containing stereotypical eye movements were retained, as eye artifacts have a broad power spectrum and do not affect narrow-band steady-state responses (Srinivasan & Petrovic, 2006).

As in our previous study (Batterink & Paller, 2017a, b), we quantified neural entrainment by measuring inter-trial coherence (ITC). ITC, also known as phase-locking value, is a measure of event-related phase locking. ITC values range from 0, indicating purely non-phase- locked activity, to 1, indicating strictly phase-locked activity. A significant ITC indicates that the EEG activity in single trials is phase-locked at a given time and frequency, rather than phase-random with respect to the time-locking experimental event. ITC was computed using a continuous Morlet wavelet transformation from .2 to 5.2 Hz via the newtimef function of EEGLAB. Wavelet transformations were computed in .0278 Hz steps with 1 cycle at the lowest frequency (.2 Hz) and increasing by a scaling factor of .5, reaching 11.75 cycles at the highest frequency (5.2 Hz). This approach was selected to optimize the tradeoff between temporal resolution at lower frequencies and frequency resolution at high frequencies (Delorme & Makeig, 2004).

Consistent with our prior results (Batterink & Paller, 2017a, b), the pattern of neural entrainment across participants was characterized by clear peaks at the word and syllable frequencies. To quantify patterns of neural entrainment for each participant, ITC values were extracted at these "peak" frequency bins, corresponding to the word and syllabic frequencies (referred to as ITC_{word} and $ITC_{syllable}$), as specified by the syllabic presentation rate used for that individual (range across participants = 260 msec–300 msec/syllable, as described under Procedure). As in our prior study, these values were averaged across each 10.8 sec epoch and across the 6 centro-frontal midline electrodes where ITC at the word and syllable frequencies showed the strongest values (FC1, C1, FCz, Cz, FC2, and C2).

To characterize the time course of learning, we used a sliding-window analysis to provide a relatively fine-grained measure of ITC changes at the word and syllable frequencies across the 12-min exposure period. After artifact correction and removal of noisy data, we grouped every 12 consecutive epochs together into "bundles" (i.e., epochs 1–12, 13–24, 25–36, etc.). We then computed ITC at the word and syllables frequencies within each of these bundles. This provided a fine-grained (bundle-by-bundle) measure of neural entrainment for each participant throughout the exposure period. However, because each bundle consists of only 12 epochs, the resulting time course data were relatively noisy. To reduce the influence of random fluctuations, the data were smoothed by using a moving average filter with a span of 5 data points (i.e., each *n*th datapoint was averaged with datapoints *n*-2, *n*-1, *n*+1, and *n*+2). Because values for the first and final bundle for each participant could not be smoothed, they were excluded from further analysis. The remaining smoothed values were used for all statistical analyses.

Robust linear mixed-effects modeling was used to test the hypothesis that ITC_{word} should increase and $ITC_{syllable}$ should decrease as a function of exposure. The smoothed ITC_{word} and $ITC_{syllable}$ values were extracted for each bundle and each participant and classified according to the following factors: participant, attention condition (full, divided), and the word presentation number corresponding the first word of the bundle (i.e., 1–774). ITC_{word} and $ITC_{syllable}$ were modeled separately, with fixed effects consisting of word presentation, attention condition, and the interaction between word presentation and attention condition, and participant included as a random intercept. Attention condition was modeled as a

categorical variable and word presentation was modeled as a continuous predictor. We hypothesized that word presentation should positively predict ITC_{word} and negatively predict $ITC_{syllable}$, reflecting a relative increase in neural entrainment to the word structure as a function of exposure to the speech stream and replicating our previous finding (Batterink & Paller, 2017a, b).

In addition, we also tested whether there was a significant interaction between attention condition and word presentation for ITC_{word}, which would provide evidence that diverting attention away from the speech stream delays statistical learning.

2.5.2. *Predicting behavioral outcome measures of statistical learning through neural entrainment measures*

As in our previous study (Batterink & Paller, 2017a, b), we quantified relative neural entrainment to the underlying words in the speech stream by computing the *Word Learning Index* (WLI), defined as the ratio of ITC at the word frequency relative to ITC at the syllable frequency (i.e., $ITC_{word}/ITC_{syllable}$). Higher WLI values indicated greater neural entrainment to the word frequency relative to the raw syllable frequency, indicative of better statistical learning. For each participant, we computed average ITC_{word} and $ITC_{syllable}$ for the entire exposure period by averaging across individual bundle values. These averaged ITC_{word} and $ITC_{syllable}$ values were then used to compute each participant's WLI.

We examined whether neural entrainment to the underlying component words of the speech stream, as indexed by the WLI, predicted behavioral performance on the familiarity-rating and target detection tasks. Our main behavioral measure of interest was the RT prediction effect on the target detection task, as we previously found this measure to be the most sensitive to statistical learning and to show the strongest associations with patterns of neural entrainment (Batterink & Paller, 2017a, b). However, we also examined whether neural entrainment patterns predicted the familiarity rating score on the recognition task. For both of these dependent variables, we tested a linear regression model that included WLI and Group as independent predictors. In addition to WLI, which represents our measure of online word segmentation, Group was also included as a predictor in order to examine whether the attentional manipulation per se predicted the RT learning effect, above and beyond effects of neural entrainment to the underlying word structure.

2.5.3. *ERP analyses*

As checks of our experimental manipulation of attention, we computed ERPs to the kaleidoscope images from the 3-back task and to the individual syllables in the speech stream.

First, we computed ERPs to the images from the 3-back task, using P300 amplitude as an index of attentional allocation to the visual task (cf. Polich, 2007; Soltani & Knight, 2000). We expected that participants in the Divided Attention group should show a robust P300 response to these images, indicating that they were attending and processing them deeply in order to perform the 3-back task. In contrast, we expected that participants in the Full Attention group should show little to no P300 response to these images, indicating that they were not attending to them as fully. After band-pass filtering from .1 to 30 Hz, epochs time-locked to each image onset were extracted from −200 to 1200 msec. Independent component analysis was used to remove eye artifacts, using the same procedure as in our previously published studies (e.g., Batterink et al., 2015a; 2015b). Using a 200-msec prestimulus baseline, mean amplitudes from 300 to 600 msec post-stimulus were computed. This time-window was selected on the basis of previous studies and on visual inspection of the data (Polich, 2007).

We also computed ERPs to the individual syllables in the speech stream, using the N100 response as an index of selective attention to the speech signal (e.g., Hillyard, Vogel, & Luck, 1998; Sanders et al., 2002). We hypothesized that learners in the Full Attention group should show a larger N100 response to individual syllables relative to participants in the Divided Attention group, reflecting an enhanced early sensory response to the input. For this analysis, we extracted epochs time-locked to each syllable onset from 0 to 300 msec from the same artifact-rejected data that was used in our neural entrainment analyses. Using a 0–10 msec baseline, mean amplitudes from 20 to 80 msec post-stimulus were computed, capturing the early sensory response to each syllable. Although the continuous nature of the speech task makes it impossible to compute a completely non-biased baseline, we selected the time interval from 0 to 10 msec as the most appropriate choice to visualize the early sensory ERP response to each individual syllable.

ERP analyses followed our usual procedures (e.g., Batterink et al., 2015a; 2015b) as follows. Amplitudes were averaged across neighboring electrodes to form nine electrode regions of interest (left anterior region: AF7, AF3, F7, F5, F3; left central region: FT7, FC5, FC3, T7, C5, C3; left posterior region: TP7, CP5, CP3, P7, P5, P3, PO7, PO3; midline anterior region: AFZ, F1, FZ, F2; midline central region: FC1, FCZ, FC2, C1, CZ, C2; midline posterior region: CP1, CPZ, CP2, P1, PZ, P2, POZ; right anterior region: AF4, AF8, F4, F6, F8; Right central region: FC4, FC6, FT8, C4, C6, T8; right posterior region: CP4, CP6, TP8, P4, P6, P8, PO4, PO8). For analysis of the P300, which showed a widespread distribution, mean amplitude values of these nine electrode regions were submitted to a repeated-measures ANOVA, with anterior–posterior axis (anterior, central, posterior) and left/right (left, midline, right) as within-participants factors, and with group (full, divided) as a between-participants factor. For analysis of the N100, which is typically maximal over frontocentral regions of the scalp, only anterior and central regions were included in the analysis. To examine whether the N100 amplitude was modulated by segmentation, we also directly compared N100 amplitude to first versus third syllables, by including syllable position (1, 3) as an additional within-subjects factor. Greenhouse–Geisser corrections were applied for factors with more than two levels.

3. Results

3.1. *Behavioral results*

3.1.1. *Three-back task (divided attention group only)*

Consistent with the expected difficulty level of the 3-back task, participants detected 48.3% (SD = 18.1%) of 3-back match trials. The false alarm rate was 28.7% (16.0%)

3.1.2. *Familiarity-rating task*

The average ratings provided for words, part-words, and nonwords are shown in Fig. 2A.

Across participants, words were rated as the most familiar, part-words intermediate, and nonwords as the least familiar, providing behavioral evidence of statistical learning [Word Category: F (2,88) = 31.4, $p < .001$; linear effect of category: F (1,44) = 66.3, $p < .001$]. In contrast to our prediction that reducing learners' attention to the speech stream would reduce explicit knowledge of the component words of the speech stream, the profile of familiarity ratings across the three word categories did not significantly differ between the Full Attention and Divided Attention groups [Attention Group x Word Category: F (2,88) = .90, $p = .41$]. The familiarity rating score was numerically higher in Full Attention than Divided Attention participants, but this group difference was not significant [Full Attention: mean = .79, SEM = .56; Divided Attention: mean = .55, SEM = .60; t (44) = 1.38, $p = .18$].

Follow-up analyses confirmed that both groups showed significant above-chance performance on this measure [Full Attention group: Word Category: F (2,40) = 20.2, $p < .001$; linear effect of category: F (1,20) = 32.6, $p < .001$; Divided Attention group: Word Category: F (2,48) = 11.7, $p < .001$; linear effect of category: F (1,24) = 32.7, $p < .001$]. Thus, both groups acquired significant explicit knowledge of the statistical structure of the speech stream. However, contrary to our hypothesis, there was no significant evidence that learners in the Divided attention group showed reduced explicit knowledge of the component words, a null finding that we consider further in the Discussion section.

3.1.3. *Target detection task*

RTs are plotted in Fig. 2B. Across all participants, RTs showed the expected decrease for later syllable positions [Position effect: F (2,88) = 44.8, $p < .001$; linear effect: F (1,44) = 75.2, $p < .001$; Position (linear effect) x Group effect: F (1,44) = 4.03, $p = .051$; $\eta^2 = .084$; $BF_{10} = 1.31$; Fig. 2B, top panel]. Across all syllable positions, participants in the Full Attention condition responded significantly more quickly than participants in the Divided Attention condition [Group effect: F (1, 44) = 4.86, $p = .033$]. Adjusting for baseline differences in RTs, the Full Attention group showed a significantly larger RT prediction effect compared to the Divided Attention group, indicative of stronger statistical learning [t (44) = −2.16, $p = .038$; Full Attention group: 18.3% decrease in RT from Position 1 to 3; Divided Attention group: 10.9% decrease; Fig. 2B, bottom panel].

Overall, participants detected 83.0% (SE = 2.6%) of targets and made an average of 18.9 (SE = 2.1) false alarms. This represents adequate performance for a task of moderate difficulty. Accuracy did not differ as a function of syllable position ($p > .7$).

3.1.4. *Correlation between familiarity-rating and target detection tasks*

The familiarity rating score and the RT prediction effect showed a significant across-participant correlation (r = .43, $p = .003$).

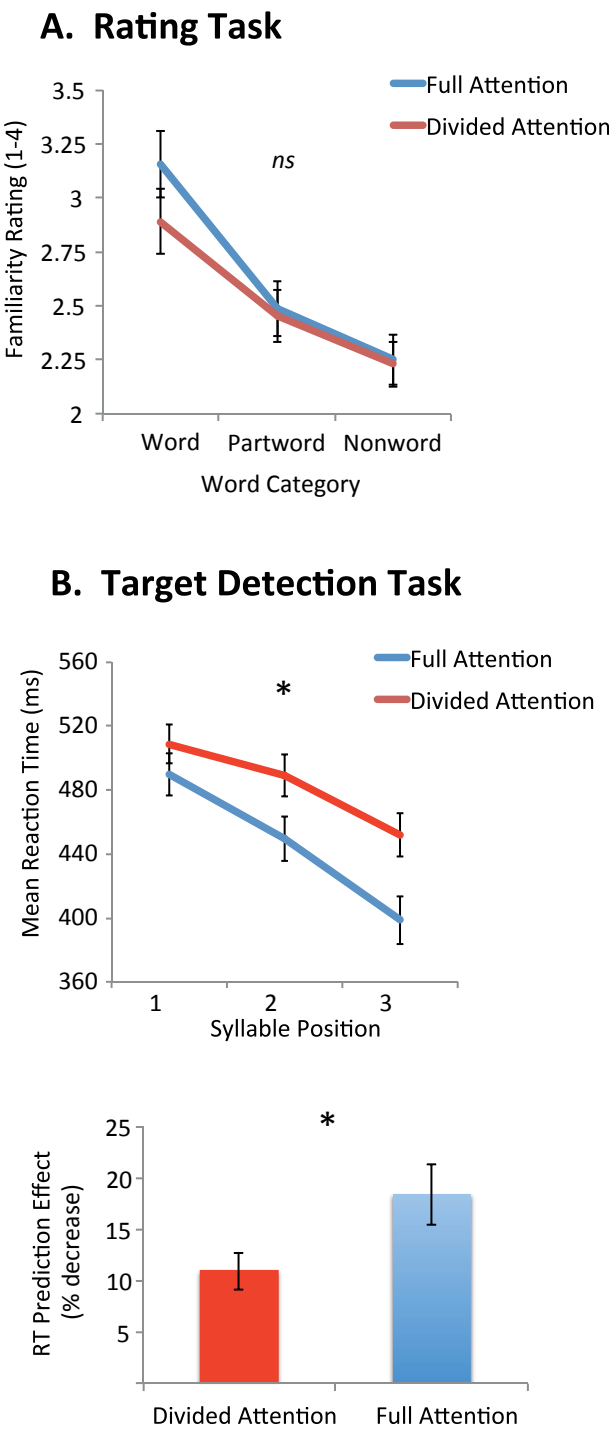

Fig. 2 – Behavioral results reflecting statistical learning, by group. Error bars represent SEM. (A) Performance on the familiarity-rating task. Participants' ratings did not differ significantly between the two groups. (B) Performance on the target detection task. Participants in the Divided Attention group responded more slowly (top panel) and showed a smaller RT prediction effect (bottom panel), relative to participants in the Full Attention group.

3.2. Neural entrainment results

3.2.1. Effect of attentional manipulation on neural entrainment

ITC as a function of frequency and group, computed across the entire exposure period, is shown in Fig. 3. Consistent with our prior results (Batterink & Paller, 2017a, b), there were clear peaks at the word and syllable frequencies, as well as the frequency corresponding to the second harmonic of the word. ITC_{word} was not significantly different between the two groups [Word: F (1,72) = .067, p = .80]. In contrast, $ITC_{syllable}$ was significantly larger in the Full Attention group [F (1,54) = 5.96, p = .018], suggesting that neural sensory responses to the individual syllables in the stream were enhanced with greater attention.

3.2.2. Time course of learning

Our time-course analysis replicated our previous findings (Batterink & Paller, 2017a, b), revealing a significant linear increase in neural entrainment at the word frequency and a significant linear decrease in neural entrainment at the syllable frequency as exposure progressed [Word Presentation: ITC_{word}: F (1,2481) = 9.89, p = .002; $ITC_{syllable}$: F (1,2478) = 6.74, p = .009]. This relative shift in entrainment towards the word frequency and away from the syllable frequency reflects the progression of learning over time. Although these effects were significant when computed across the entire exposure period, a closer examination of the data revealed that changes in neural entrainment as a function of word repetition primarily occurred during the first half of exposure [ITC_{word}: F (1,1282) = 21.5, p < .001; $ITC_{syllable}$: F (1,1278) = 32.8, p < .001; Fig. 4], with no further changes occurring in the second half of exposure [ITC_{word}: F (1,1154) = 1.83, p = .18; $ITC_{syllable}$: F (1,1153) = .19, p = .67]. Thus, learning progressed in the first half of exposure only, and was followed by a plateau in the second half of exposure, with ITC values in the second half reflecting final levels of attainment.

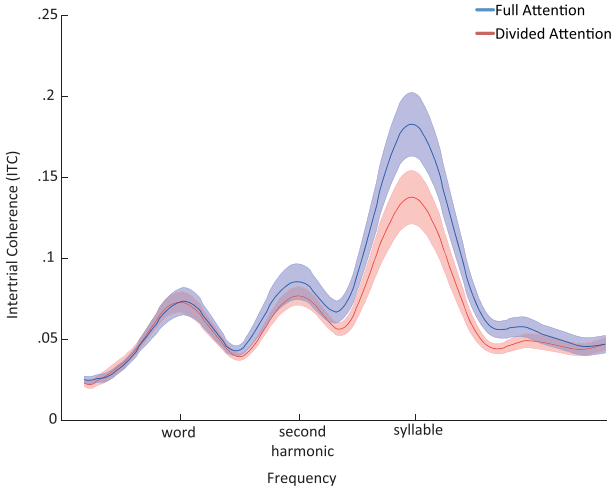

Fig. 3 – ITC as a function of frequency and group, computed across the entire exposure period. Shaded regions represent the mean ± SEM. Slightly different stimulus presentation rates were used across participants (see text), and thus the frequencies depicted on the x-axis are expressed relative to the word, second harmonic, and syllabic frequencies used for each individual participant, rather than in numerical terms.

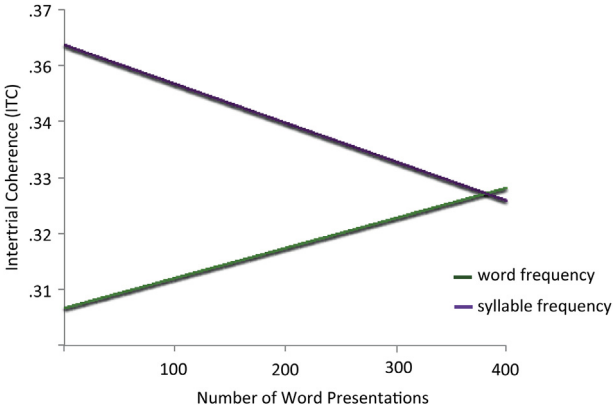

Fig. 4 – Modeled progression of ITC at the word and syllable frequency as a function of exposure, based on parameter estimates of fixed effects in the linear mixed-effects model within the first half of exposure. ITC at the word level increased significantly as a function of exposure, while ITC at the syllabic level decreased significantly as a function of exposure, reflecting online learning.

To examine whether learning showed a different progression over time as a function of our attentional manipulation, we compared the time course of learning between the two groups within the first half of exposure, when there was evidence of a learning progression across participants. Change in ITC_{word} over the first half of exposure did not significantly differ between the two groups [Attention condition x Word Presentation: F (1,1282) = .30, p = .59]. Final ITC_{word} values, computed over the second half of exposure, also did not differ between the two groups [Attention condition: F (1,543) = .89, p = .35]. In contrast, $ITC_{syllable}$ showed a more rapid decline over the first half of exposure in the Full Attention condition compared to the Divided Attention condition [Attention condition x Word Presentation: F (1,1153) = 14.6, p < .001]. However, this finding is somewhat difficult to interpret, as the Full Attention group had higher overall $ITC_{syllable}$ values. Although $ITC_{syllable}$ declined more rapidly in the Full Attention group in the first half of exposure, these values remained higher in the Full Attention group compared to the Divided Attention group in the second half of exposure [Attention condition: F (1,255) = 11.4, p = .001].

In summary, no significant differences were found in the time course of online word perception between the two groups, as reflected by neural entrainment at the word frequency. That is, both groups appear to have extracted the word structure from the speech stream at similar rates. Interestingly, we also found evidence that learning progressed primarily in the first half of the exposure period (~6 min) for both groups, reaching a plateau by the second half of exposure.

3.2.3. *Relationship between neural entrainment and behavioral measures of statistical learning*

Relative neural entrainment to the word frequency across the Exposure period significantly predicted the RT prediction effect on the target detection task, our main measure of statistical learning. A regression model with WLI and Group as predictors significantly predicted the RT prediction effect [$F(2,43) = 5.40$, $p = .008$]. Both variables made significant independent contributions to the model [Group: $t(43) = 2.84$, $p = .007$; WLI: $t(43) = 2.31$, $p = .026$]. These results indicate that the Full Attention group showed a larger RT prediction effect than the Divided Attention group. In addition, independently of this group effect, learners who showed relatively greater neural entrainment at the word frequency compared to the syllable frequency, as reflected by a larger WLI, showed a larger RT prediction effect, indicative of better statistical learning (Fig. 5).

In contrast to performance on the target detection task, the regression model with WLI and Group did not predict the familiarity rating score [$F(2,43) = 1.22$, $p = .31$]. Neither predictor in the model was significant (both p values > .14). Performance on the target detection task may be a more sensitive measure of statistical learning compared to performance on the familiarity-rating task. This finding is consistent with results from our previous study, in which we also found no significant relationship between neural entrainment patterns and explicit familiarity rating scores (Batterink & Paller, 2017a, b).

3.3. *ERP manipulation checks*

3.3.1. *P300 response to images from 3-back task*

Participants in the Full attention condition were asked to attend fully to the speech stream, whereas participants in the Divided Attention condition were asked to concentrate on the kaleidoscope images and ignore the speech stream. Reflecting this experimental manipulation, the P300 response to visual images for participants in the Divided Attention condition was much larger than for participants in the Full Attention condition [Group effect: $F(1,44) = 61.74$, $p < .001$; Fig. 6A]. Across all participants, the P300 was maximal over midline and posterior electrodes [Anterior/Posterior: $F(2,88) = 23.4$, $p < .001$; Left/Midline/Right: $F(2,88) = 11.2$, $p < .001$], consistent with the expected distribution of a P3b. These results serve as a manipulation check and confirm that participants in the Divided Attention condition allocated more attention to the visual images than participants in the Full Attention condition.

3.3.2. *Early N1 sensory response to individual syllables*

As shown in Fig. 6B, participants in the Divided Attention condition showed a reduced early negativity/N100 response to individual syllables in the speech stream, relative to participants in the Full Attention condition [Group effect: $F(1,44) = 4.27$, $p = .045$]. This group difference was marginally larger over frontal sites compared to central sites [Group effect x Anterior/Posterior: $F(1,44) = 3.25$, $p = .078$]. These results provide additional verification that participants in the Divided Attention condition allocated less attention to the auditory speech stream than did participants in the Full Attention condition. Although N100 amplitude was numerically larger to first syllables compared to third syllables, this difference was very small and not statistically significant [Position effect: $F(1,44) = .57$, $p = .45$; all interactions *ns*, $p > .2$; Fig. 6C].

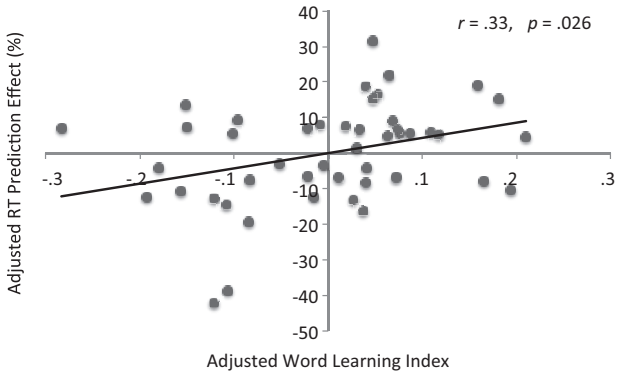

Fig. 5 – Partial correlation between the online Word Learning Index (WLI) and the RT prediction effect on the target detection task, controlling for the effect of group (Full *vs* Divided Attention). The adjusted values shown in this graph represent the residual values for the WLI and RT effect variables after adjusting for the effect of attention group. Greater neural entrainment at the word level relative to the syllable level predicts a larger RT prediction effect.

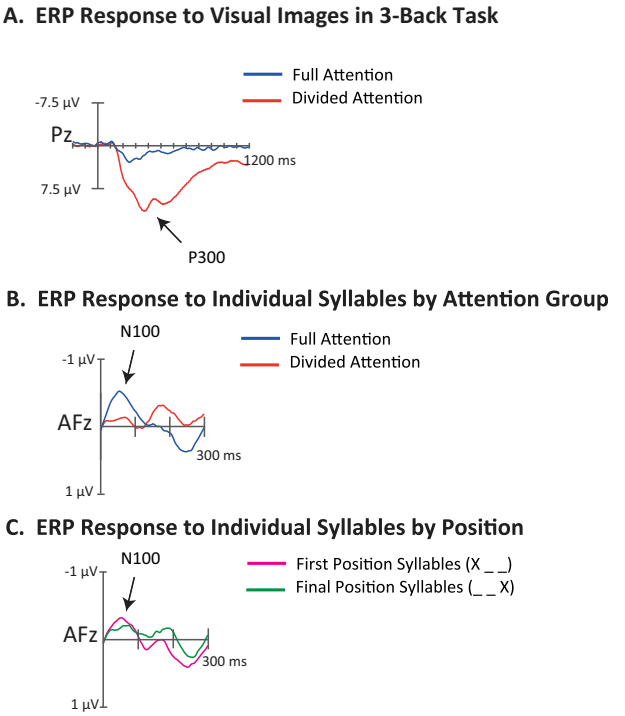

Fig. 6 – ERP results to visual and auditory stimuli, by group. (A) ERP grand averages to images from visual 3-back task. (B) ERP grand averages to individual syllables in the speech stream. (C) ERP grand averages to individual syllables in the speech stream as a function of syllable position, across all participants.

4. Discussion

We designed this experiment to test whether concentration of voluntary, top-down attentional resources influences two dissociable components of statistical learning, namely the perceptual binding of neighboring syllables into component words, and subsequent memory storage and retrieval of the acquired word representations. Both of these components are expected to critically influence performance on offline tests of statistical learning. Focused attention to the speech stream influenced the sensory processing of stimuli as expected, but interestingly did *not* impact the online perceptual binding of syllables. Learners in both the Divided and Full Attention groups showed similar overall levels of neural entrainment to the underlying words, and also showed similar learning progressions, as reflected by time-course analyses. In contrast, attention did impact the acquired memory representations, at least as assessed by performance on one of the two offline tasks. Specifically, reducing attention to the speech stream resulted in smaller RT prediction effects on the target detection task, potentially reflecting weaker representations of the words in memory. However, contrary to our hypothesis, no effect of our attentional manipulation on offline familiarity ratings was found. In addition, learners in the Divided Attention group still showed evidence of robust statistical knowledge on both the familiarity-rating and target detection tasks. Taken together, these findings suggest that (1) concentration of attention during training does not strongly influence the perceptual component of statistical learning, but plays some role in subsequent memory-related processes, and (2) both the perceptual and memory-related components of statistical learning can occur "in the background," while learners are engaged in a demanding concurrent task, at least to some degree.

We confirmed using neural measures that our attentional manipulation was successful and that sensory responses to the individual syllables in the stream were enhanced with greater attention. One group of learners was allowed to focus fully on the speech stream (Full Attention group), whereas the other group was required to perform a demanding visual N-back task with the auditory speech stream presented outside their primary focus of attention (Divided Attention group). Neural entrainment to the raw syllable frequency was significantly larger in the Full Attention group, reflecting enhanced sensory processing of individual syllables. In addition, ERP measures demonstrated that participants in the Divided Attention group allocated greater attention/limited-capacity processing resources to the visual images, and correspondingly less of these resources to the auditory speech stream, relative to participants in the Full Attention group. First, P300 to the visual images was much larger in the Divided Attention group compared to the Full Attention group. Given that posterior P300 is associated with voluntary attention, task-relevant processing, context updating, and memory storage (Polich, 2007; Soltani & Knight, 2000), this finding indicates that Divided Attention participants allocated more attentional resources to the visual task relative to the Full Attention participants. In addition, participants in the Full Attention group showed an enhanced early N1 response to the individual syllables. In light of previous ERP studies showing modulations by selective auditory attention (e.g., Hillyard, Hink, Schwent, & Picton, 1973), this N1 result demonstrates that Full Attention participants allocated greater processing resources to the speech stream relative to Divided Attention participants, and converges with the observed neural entrainment effect at the syllable frequency.

Overall, across both groups, our results replicated our previous findings on the time course of neural entrainment and the relation of neural entrainment to other measures of statistical learning (Batterink & Paller, 2017a, b). We again demonstrated that neural entrainment simultaneously increases at the word level and decreases at the syllable level as a function of exposure, tracking the progression of learning. This frequency shift in entrainment over time points to a gradual perceptual transformation that accompanies statistical learning, whereby processing is initially dominated by low-level, bottom-up sensory input at the single syllable level but is increasingly driven by the perception of larger integrated words. That is, as syllables are bound together into the most relevant units (in this case, trisyllabic words), entrainment is reduced at nontarget or irrelevant frequencies while enhanced at the most relevant item frequency. Interestingly, by using a fine-grained analysis of the online measure to characterize the time course of learning, we found that learning progressed only in the first half of the exposure period (~6 min), with a plateau occurring in the second half. This result is consistent with recent RT evidence mapping the trajectory of statistical learning, which points to a relatively fast increase in learning early in exposure, followed by a period of no further learning that consists of only random fluctuations around a fixed value (Siegelman et al., 2017). These time course data also mirror ERP effects indexing statistical learning, such as the N400 effect, which emerge rapidly after only a short period of exposure to structured input and then stabilize (Cunillera et al., 2009) or even disappear (Abla, Katahira, & Okanoya, 2008). We also replicated our finding that overall level of neural entrainment to words relative to syllables predicts performance on the target detection task. Learners who showed a greater tendency to bind neighboring syllables into coherent word units, as measured through neural entrainment, also showed a larger RT effect, reflecting greater facilitation in processing from knowledge acquired through statistical learning. Taken together with our prior report (Batterink & Paller, 2017a, b), these findings further validate our neural entrainment measure of statistical learning.

4.1. *Effects of attention on dissociable components of statistical learning*

As described in the Introduction, the goal of our study was to investigate whether attention and associated limited-capacity resources play a role in two conceptually dissociable aspects of statistical learning: online perceptual binding of underlying words and subsequent memory storage/retrieval processes. Our results suggest that focused attention does not play a major role in the online perceptual component of statistical learning as assessed through our neural entrainment measure. Although entrainment at the syllable level was

enhanced with greater attention—presumably reflecting a gain in sensory processing of the individual syllables—no group differences were observed at the word level. In addition, we found no evidence that the two groups differed in the progression of neural entrainment to words, even using a fine-grained approach to characterize the time course of learning. These results suggest that reducing attentional processes to the speech stream does not greatly impact learners' ability to extract relevant statistical probabilities from speech and perceptually group individual syllables into component words.

In contrast to our finding that attentional resources do not impact neural entrainment to underlying component words, the two groups showed significant differences on one of our two offline memory tasks, the target detection task. Even after accounting for baseline differences in RT, learners in the Full Attention condition showed a larger RT prediction effect, as assessed by the relative speed-up to predictable compared to unpredictable syllables. This result indicates that allocating greater attention to the speech stream resulted in stronger memory representations of the component words, allowing Full Attention learners to more efficiently process and predict syllables that occurred in later, more predictable positions. In addition, learners in the Full Attention condition responded significantly faster to all target syllables (across the three syllable positions) compared to learners in the Divided Attention condition. One possibility is that stronger acquired memory representations of the component words may have allowed Full Attention learners to better predict later syllable targets as well as initial syllable targets, which are still highly predictable given the limited word inventory of the speech stream (consisting of a .33 transitional probability, given that no word was repeated immediately within the stream). Another possibility is that this overall RT difference between the groups may reflect a difference in perceptual sharpening or perceptual tuning produced by the attentional manipulation. By attending more fully to the speech stream, learners in the Full Attention condition may have developed clearer and sharper perceptual representations of the individual syllables, which may then have allowed them to respond more quickly to all syllables, regardless of their position within the component words.

However, on the familiarity-rating task, we found no significant differences in accuracy between the two groups, with both groups showing evidence of learning. Although learners in the Full attention group showed better accuracy numerically on this task than learners in the Divided attention group, this difference was small and did not reach significance ($p = .18$). This finding is counter to our original hypothesis that the Divided attention group would show impaired familiarity-rating accuracy, consistent with the well-established finding that explicit memory is impaired when attention is divided between two tasks at encoding (e.g., Craik, Govoni, Naveh-Benjamin, & Anderson, 1996). One possible reason for this null group effect may be a lack of sensitivity of the familiarity-rating task. We and others have shown in prior work that direct memory tasks, such as recognition tests and explicit familiarity ratings, may be less sensitive indices of statistical learning compared to indirect memory tasks such as the target detection task (Batterink et al., 2015a, b; Siegelman et al., 2016; Siegelman et al., 2017). In particular, direct measures may underestimate the total amount of knowledge produced by statistical learning and can be contaminated by other cognitive processes not of direct interest, such as individual differences in memory retrieval or strategic processing. Consistent with this idea, participants' performance on this task across both groups was relatively poor, with the average familiarity rating for words versus nonwords differing by only .78 out of a maximum possible difference of 3 (see Fig. 2A). Thus, although the task was sufficiently sensitive to reveal evidence of statistical learning across learners, it may not have been sensitive enough to demonstrate group differences in memory, at least with the current sample size. We speculate that a more sensitive measure of explicit memory may have successfully revealed explicit memory differences between the two groups, though testing this idea awaits further methodological refinement.

The present study has many parallels in design and aims to a recent study of rule learning (López-Barroso, Cucurell, Rodríguez-Fornells, & de Diego-Balaguer, 2016). This study examined the role of attention in the learning of nonadjacent AXC rule dependencies, in which the third syllable of a nonsense word is predicted by the first syllable. Attention was manipulated through a word monitoring task in which participants were asked to detect *C* targets contained within only one of three *AXC* structures; in other words, A_1XC_1 was attended, while A_2XC_2 and A_3XC_3 rules were unattended. Learning was assessed using both an indirect target detection task as well as a direct test requiring explicit judgments of rule violations. Similar to the present study, attention resulted in faster overall responses on the target detection task, with participants responding faster to previously attended targets (i.e., C_1) compared to previously unattended targets (i.e., C_2 and C_3). Nonetheless, evidence of learning on this target detection task was also found for unattended rules, as assessed by faster responses to rules (e.g., A_2XC_2) compared to non-rules (e.g., XXC_2). Consistent with our findings, this suggest that increased attention enhances memory-related processes and facilitates performance on subsequent indirect tests of memory, but that some learning can also occur even with minimal focused attention.

To summarize, we found evidence that our attentional manipulation influenced performance on the target detection task, an indirect measure of memory, but not our online neural entrainment measure or on familiarity ratings. Given our assumption that the post-exposure target detection task is sensitive to both the perceptual binding and subsequent memory storage components of statistical learning, whereas neural entrainment is sensitive to only perceptual binding, these findings suggest that attention impacts only the memory storage component of statistical learning. We found no evidence that focused attention modulates perceptual binding.

4.2. *Statistical learning can proceed outside the focus of attention*

Our study also sheds light on the extent to which statistical learning can proceed outside the focus of attention. Interestingly, learners in the Divided Attention condition showed

robust evidence of learning on *all* measures of statistical learning, including the WLI, offline familiarity ratings, and RTs on the target detection task. Taken together, these results indicate that both components of statistical learning—perceptual binding of underlying words and subsequent memory-related processes—can proceed while learners are performing an attention-demanding visual task. At the same time, we cannot rule out the possibility that *some* amount of voluntary attention and working memory resources are necessary for both components of statistical learning. Although they had no reason to do so intentionally, learners in the Divided Attention condition may have alternated their focus of attention and associated processing resources from the visual task to the auditory speech stream. These residual levels of attention and working memory resources may have been sufficient to support learning. Thus, our study is constrained by the same limitation facing virtually all studies of selective attention—we can experimentally reduce attentional resources to a given process of interest, but cannot completely abolish it. Nonetheless, our converging results from all measures conclusively demonstrate that statistical learning can occur with stimuli that learners have been instructed to ignore, outside of the direct focus of attention.

The current findings are consistent with several previous studies that indicate that learners can carry out attention-demanding tasks while simultaneously acquiring the statistics of sensory input that lies outside their primary focus of attention (Fernandes et al., 2010; Musz et al., 2015; Saffran et al., 1997), at least for salient statistical units (Fernandes et al., 2010). However, on the surface they are inconsistent with work conducted by Toro et al. (2005) and Palmer and Mattys (2016), which both involved visual *N*-back tasks to divert learners' attention away from the speech stream, similar to our experimental manipulation. In contrast to the current findings, both these studies found that reducing attention to the speech stream negatively impacted performance on a forced-choice recognition task. One critical feature that may distinguish our study from these prior studies is the stimulus presentation rate used in the visual distractor task, which may have placed heavier demands on selective attention and weaker demands on working memory relative to the present study. In the Toro study (Experiment 2), participants performed a 1-back task on a stream of pictures, presented at a rate of either 500 or 750 msec per item. Palmer and Mattys (2016) used a 2-back task on unnameable visual shapes presented at a rate of 750 msec. In contrast, we used a 3-back task, with each visual stimulus presented for relatively long durations (2400–5000 msec). Thus, relative to the distractor tasks used by these previous studies, our task was less taxing at the perceptual level of attentional selection, but placed higher loads on working memory, requiring participants to actively maintain multiple visual stimuli for long periods of time.

Taken together, these results suggest that statistical learning may critically depend on the attentional selection of input at the early perceptual stage, but to a lesser extent on later, post-perceptual resources, including working memory. In other words, as long as initial perceptual encoding of the incoming stimuli occurs without impairment, statistical learning may occur even when central executive functions and limited-capacity resources are occupied through a demanding secondary task.

4.3. *Potential implications for language acquisition*

One of the major findings of this study—that statistical learning can occur even when attentional and limited-capacity central resources are being actively consumed by a concurrent task—has important implications for language acquisition. This result suggests that learners may benefit from exposure to a novel spoken language even when their attentional focus is engaged by a competing task, such that expertise with statistical regularities of the input could be acquired without specific intention or mental effort. That is, acquiring the statistics of linguistic input may occur not only when learners have the opportunity to actively focus on the target language, but also "in the background." By capitalizing on this feature of statistical learning, learners may be able to speed up the process of speech segmentation of an unfamiliar language. For example, immigrants to a new country may facilitate their language acquisition by regularly having the radio or TV on in their home while they go about everyday tasks. By extension, exposure to the statistics of a new language may even be helpful when presented during sleep (Batterink & Paller, 2017b; Schreiner & Rasch, 2017). Although it will be important to determine whether the artificial speech segmentation paradigm used in this study scales up to language learning outside the laboratory, the finding that statistical learning can occur without focused attention may potentially allow for novel, low-cost, and low-effort interventions to enhance some aspects of language acquisition, particularly in adult learners for whom acquiring a new language can be particularly challenging.

4.4. *Future directions and conclusions*

A number of questions concerning the role of attention and limited-capacity resources in statistical learning remain. Although the perceptual component of statistical learning was not impacted by our attentional manipulation using a cross-modality distractor task, attentional costs may be demonstrated under other circumstances, as we have noted previously. For example, given that effects of attention are typically more robust within modalities than across modalities (i.e., Duncan, Martens, & Ward, 1997; Soto-Faraco & Spence, 2002; Treisman & Davies, 1973; Wickens, 1984), a question for future research is whether learning compromised when attention is diverted to a competing auditory task (e.g., detecting auditory targets in a competing auditory stream). Another question is the extent to which competing linguistic tasks, such as silent reading, may compromise this type of learning.

Taken together, the present results suggest that focused, directed attention to input plays a limited and non-essential role in a critical component of statistical learning, the online perceptual binding of syllable forming component words. In contrast, reduced attention significantly impacts the memory storage component of statistical learning. Nonetheless,

evidence of robust learning was found at both levels of attention, across all experimental measures, indicating that both components of statistical learning can occur even when speech input is outside of learners' targeted focus of attention. In addition to providing insight into the necessary and sufficient conditions for statistical learning, these results also have important implications for language acquisition, potentially opening up previously overlooked opportunities for learning.

Acknowledgments

We would like to thank Kelsey Aaronson for her help in data collection and two anonymous reviewers for their helpful comments. We are also grateful for funding from the National Institute of Child Health and Human Development (F32 HD078223 to L. Batterink), the National Institute of Neurological Disorders and Stroke (T32 NS047987) and the National Science Foundation grant BCS-1461088.

REFERENCES

Abla, D., Katahira, K., & Okanoya, K. (2008). On-line assessment of statistical learning by event-related potentials. *Journal of Cognitive Neuroscience, 20*(6), 952–964.

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex, 90*, 31–45.

Batterink, L. J., & Paller, K. A. (2017). Sleep-based memory processing facilitates grammatical generalization: Evidence from targeted memory reactivation. *Brain and Language, 167*, 83–93.

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.

Batterink, L. J., Reber, P. J., & Paller, K. A. (2015). Functional differences between statistical learning with and without explicit training. *Learning and Memory, 22*, 544–556.

Buiatti, M., Pena, M., & Dehaenelambertz, G. (2009). Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *Neuroimage, 44*(2), 509–519.

Craik, F. I., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology General, 125*, 159–180.

Cunillera, T., Càmara, E., Toro, J. M., Marco-Pallares, J., Sebastián-Galles, N., Ortiz, H., et al. (2009). Time course and functional neuroanatomy of speech segmentation in adults. *Neuroimage, 48*(3), 541–553.

Cunillera, T., Toro, J. M., Sebastián-Gallés, N., & Rodríguez-Fornells, A. (2006). The effects of stress and statistical cues on continuous speech segmentation: An event-related brain potential study. *Brain Research, 1123*(1), 168–178.

De Diego Balaguer, R., Toro, J. M., Rodriguez-Fornells, A., & Bachoud-Lévi, A.-C. (2007). Different neurophysiological mechanisms underlying word and rule extraction from speech. *PLoS One, 2*(11), e1175.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*, 9e21.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience, 19*, 158–168.

Duncan, J., Martens, S., & Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature, 387*, 808–810.

Fernandes, T., Kolinsky, R., & Ventura, P. (2010). The impact of attention load on the use of statistical information and coarticulation as speech segmentation cues. *Attention Perception and Psychophysics, 72*, 1522–1532.

Fougnie, D. (2008). The relationship between attention and working memory. In Noah B. Johansen (Ed.), *New research on short-term memory*. Nova Science Publishers, Inc.

Franco, A., Eberlen, J., Destrebecqz, A., Cleeremans, A., & Bertels, J. (2015). Rapid serial auditory presentation: A new measure of statistical learning in speech segmentation. *Experimental Psychology, 62*, 346–351.

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science, 182*, 177–180.

Hillyard, S. A., Vogel, E. K., & Luck, S. J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: Electrophysiological and neuroimaging evidence. *Philosophical Transactions of the Royal Society B, 353*, 1257–1270.

Kabdebon, C., Pena, M., Buiatti, M., & Dehaene-Lambertz, G. (2015). Electrophysiological evidence of statistical learning of long-distance dependencies in 8-month-old preterm and full-term infants. *Brain and Language, 148*, 25–36.

Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: Is it long-term and implicit? *Neuroscience Letters, 461*, 145–149.

Lehiste, I. (1960). An acoustic phonetic study of open juncture. *Phonetica Supplementum, 5*, 1–54.

López-Barroso, D., Cucurell, D., Rodríguez-Fornells, A., & de Diego-Balaguer, R. (2016). Attentional effects on rule extraction and consolidation from speech. *Cognition, 152*, 61–69.

Musz, E., Weber, M. J., & Thompson-Schill, S. L. (2015). Visual statistical learning is not reliably modulated by selective attention to isolated events. *Attention Perception and Psychophysics, 77*, 78–96.

Nozaradan, S., Peretz, I., Missal, M., & Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meat. *Journal of Neuroscience, 31*(28), 10234–10240.

Palmer, S. D., & Mattys, S. L. (2016). Speech segmentation by statistical learning is supported by domain-general processes within working memory. *The Quarterly Journal of Experimental Psychology, 69*, 2390–2401.

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*, 2128–2148.

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

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

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

Sanders, L. D., Newport, E. L., & Neville, H. J. (2002). Segmenting nonsense: An event-related potential index of perceived onsets in continuous speech. *Nature Neuroscience, 5*(7), 700–703.

Schreiner, T., & Rasch, B. (2017). The beneficial role of memory reactivation for language learning during sleep: A review. *Brain and Language, 167*, 94–105.

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2016). Towards a theory of individual differences in statistical learning. *Philosophical Transactions B, 372*, 20160059.

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2017). Redefining "learning" in statistical learning: What does an

online measure reveal about the assimilation of visual regularities? *Cognitive Science*, 1–36.
Soltani, M., & Knight, R. T. (2000). Neural origins of the P300. *Critical Reviews in Neurobiology, 14*, 199–224.
Soto-Faraco, S., & Spence, C. (2002). Modality-specific auditory and visual temporal processing deficits. *Quarterly Journal of Experimental Psychology*, 55, 23–40.
Srinivasan, R., & Petrovic, S. (2006). MEG phase follows conscious perception during binocular rivalry induced by visual stream segregation. *Cerebral Cortex, 16*, 597–608.
Toro, J. M., Sinnet, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, B25–B34.
Treisman, A. M., & Davies, A. (1973). Divided attention between ear and eye. In S. Kornblum (Ed.), *Attention and performance IV* (Vol. 4, pp. 101–117). New York: Academic Press.
Turk-Browne, N. B., Junge, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology General, 134*, 552–564.
Voss, J., & Paller, K. A. (2009). An electrophysiological signature of unconscious recognition memory. *Nature Neuroscience, 12*, 349–355.
Wickens, C. (1984). Processing resources in attention. In R. Parasuraman, & D. Davies (Eds.), *Varieties of attention* (pp. 63–102). London: Academic Press.