CORTEX 192 (2025) 242–270

Available online at www.sciencedirect.com

ScienceDirect

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

Registered Report

Neural evidence for linguistic statistical learning is independent of rhythmic and cognitive abilities in neurotypical adults

I.M. van der Wulp [a,*], M.E. Struiksma [a], L.J. Batterink [b] and F.N.K. Wijnen [a]

[a] Department of Languages, Literature and Communication, Institute for Language Sciences, Utrecht University, Utrecht, the Netherlands

[b] Department of Psychology, Western Institute for Neuroscience, Western University, London, ON, Canada

ARTICLE INFO

Article history:
Received 30 September 2025
Revised 30 September 2025
Accepted 30 September 2025
Action Editor Elizabeth Wonnacott

Keywords:
Statistical learning
Speech segmentation
Individual differences
Neural oscillations
EEG
Phase-locking
Rhythmic abilities
Cognitive abilities

ABSTRACT

Statistical Learning (SL) is an essential mechanism for speech segmentation. Individual differences in SL ability are associated with language acquisition. For instance, better SL correlated with a larger vocabulary size and impaired SL was found in populations with language impairments. The aim of the current study was to contribute to uncovering the underpinnings of individual differences in auditory SL for word segmentation. We hypothesized that individuals with better musical – specifically rhythmic – abilities would show better SL. Participants ($N = 106$) were exposed to an artificial language consisting of trisyllabic nonsense words. *Electroencephalography* (EEG) measures of neural entrainment to the auditory signal allow online assessment of SL. The current study used this method to measure individual SL performance during exposure. To assess individual differences, we linked the neural measure of SL to a battery of tests measuring rhythmic, musical, and cognitive abilities, as well as vocabulary size. We replicated earlier work, finding both online (neural) and offline (behavioral) evidence of SL in our sample. In contrast to our expectations regarding individual differences, we found evidence for the null hypothesis regarding correlations between the tests of rhythmic ability and the neural measurement of SL. Exploratory analyses concerning working memory remained inconclusive, while exploratory analyses regarding vocabulary size yielded moderate evidence for a small correlation with the neural measure of SL. Overall, our results suggest that linguistic SL is largely independent from abilities in other cognitive domains, including rhythmic processing and musical abilities, as measured within a sample of healthy, typically developed adults.

* *Corresponding author*. Department of Languages, Literature and Communication, Faculty of Humanities, Utrecht University, Trans 10, 3512 JK, Utrecht, the Netherlands.
E-mail address: i.m.vanderwulp@uu.nl (I.M. van der Wulp).

https://doi.org/10.1016/j.cortex.2025.09.012

1. Introduction

1.1. Statistical learning for speech segmentation

Individuals acquiring a new language untutored face the challenge of *speech segmentation*[1]: dividing the continuous streams of speech sounds they hear in their environment into meaningful words. This is an important (first) step in acquiring a vocabulary and it is fundamentally linked to further linguistic development (Erickson & Thiessen, 2015; Evans et al., 2009; Newman et al., 2016; Rodríguez-Fornells et al., 2009; Siegelman, 2020; Singh et al., 2012; Zhang et al., 2021).

Statistical learning (SL) is thought to support speech segmentation and refers to the process of becoming sensitive to the statistical structure of a stimulus stream (Saffran, Aslin et al., 1996; Saffran, 2003). The statistical structure useful for segmenting continuous speech can be quantified as *transitional probabilities* between neighboring syllables[2]; the probability that a syllable *X* is directly followed by a syllable *Y*, given the overall frequency of *X* (Saffran, Newport et al., 1996). In natural language, transitional probabilities are higher for syllable transitions within words than for syllable transitions spanning word boundaries (Saffran, 2003). Transitional probabilities can thus serve as a statistical cue for the learner as to where a word boundary is likely to occur.

Research assessing SL in the laboratory has found salient inter-individual differences in SL performance (e.g., Batterink & Paller, 2017; Bogaerts et al., 2022), which have been linked to individual variability in language acquisition (Erickson & Thiessen, 2015; Siegelman, 2020; Singh et al., 2012). However, it is currently still unknown which factors underlie these individual differences. Therefore, the aim of the current study was to contribute to the knowledge in the field regarding the underpinnings of individual differences in auditory SL for word segmentation.

1.2. Assessing statistical learning in the laboratory

Using artificial language learning paradigms, multiple experimental studies have found that both adults and infants are able to use SL to segment 'words' (multi-syllabic sequences) from a continuous speech stream (e.g., Batterink & Paller, 2017; Choi et al., 2020; François, Chobert et al., 2012; Pinto et al., 2022; Saffran, Aslin et al., 1996; Saffran, Newport et al., 1996; Schön & François, 2011). These studies typically employ a *familiarization phase* in which participants passively listen to the stimulus stream made up of the concatenated words without any pauses or other acoustic cues to word boundaries. This phase is then followed by a *test phase* in which participants usually perform a *two-alternative forced choice* (2AFC) task. In this task, participants hear 'words' (previously presented patterns) and 'foils' (syllables presented in a recombined order) and are asked to identify the previously presented words. The rationale is that accuracy on the 2AFC task above chance level (50%) provides evidence that the participant has successfully acquired the patterns through SL.

However, the 2AFC task has often been criticized for tapping into explicit memory and meta-cognitive decision making (Bogaerts et al., 2022; François, Tillmann et al., 2012). Alternatively, other tasks have been proposed to probe SL outcomes by evaluating the expression of *implicit memory*. SL is often referred to as 'implicit learning' (Erickson & Thiessen, 2015; Perruchet & Pacton, 2006) and, when measured by implicit memory tasks, can reveal learning in the absence of explicit knowledge or awareness of the regularities (Arciuli, 2017; Batterink et al., 2015, 2019; Schön & François, 2011). One task that was designed to tap into implicit memory of statistical regularities in speech input is the target detection task (TDT; Batterink, 2017; Batterink et al., 2015; Batterink & Paller, 2017, 2019; Kim et al., 2009; Moreau et al., 2022; Turk-Browne et al., 2005). In this task, participants are presented with a target syllable and subsequently hear a shortened version of the stimuli presented during the familiarization phase. They are asked to press a button as quickly and accurately as possible when they hear the target syllable in the stimulus stream. If participants have learned the tri-syllabic words, they should show a gradual facilitation pattern expressed by faster reaction times (RTs) towards the word-final syllables, which are the most predictable compared to the second and first syllable.

Implicit measures such as the TDT are a step in the right direction for assessing SL in the laboratory. However, they are still administered *after* the familiarization phase and are thus also unable to access the learning process itself (e.g., Bogaerts et al., 2022; Schön & François, 2011). It has been proposed that SL for word segmentation is a two-step process, which starts with identification of the individual word forms – the process of segmenting the speech input – followed by long-term memory formation for these extracted word forms (Batterink & Paller, 2017; Erickson & Thiessen, 2015; Rodríguez-Fornells et al., 2009). The conventional techniques probe the second of these steps and therefore can only provide *indirect* evidence on the first step. A promising new avenue in SL research is therefore the recording of neural oscillations through *electroencephalography* (EEG) during the familiarization phase (Batterink & Paller, 2017, 2019; Choi et al., 2020; Moreau et al., 2022; Pinto et al., 2022; Zhang et al., 2021). Neural oscillations have previously been shown to *phase-lock*[3] to the rhythm of a perceived auditory stimulus such as language (Daikoku & Goswami, 2022; Giraud & Poeppel, 2012; Peelle & Davis, 2012). Batterink and Paller (2017) captured this neural entrainment to the speech streams by computing the *Inter-Trial Coherence* (ITC) to the frequencies corresponding to the presentation rate of the syllables (3.3 Hz; each syllable was presented every 300 msec) and the tri-syllabic words (1.1 Hz; 900 msec). Their results showed that there was progressively more phase-locking during exposure at the word frequency – as indicated by an increasing ITC over time – along with decreasing phase-locking at the syllable frequency in the

[1] This is also frequently referred to as *word segmentation*.

[2] Syllables are a basic unit of spoken language (e.g., Poeppel & Assaneo, 2020) and therefore transitional probability computations are made based on neighboring syllables for speech segmentation.

[3] Also: *entrain, synchronize*. The phase of the neural oscillations aligns with the phase of the input signal.

structured speech stream. From these ITC values, the authors computed a *Word Learning Index* (WLI), which provides a relative measure of sensitivity to the trisyllabic structure of the input in the structured condition:

$$WLI = \frac{ITC_{word\ frequency}}{ITC_{syllable\ frequency}}$$

Thus, the WLI increased during exposure to the structured stream. This was contrasted to a control condition comprising of a random speech stream which did not contain underlying regularities, and the WLI in this condition did not change over time. The WLI furthermore correlated significantly with individual performance on the TDT. Thus, the study by Batterink and Paller (2017), as well as subsequent experiments with the same frequency-tagging paradigm (Batterink & Paller, 2019; Choi et al., 2020; Moreau et al., 2022; Pinto et al., 2022; Zhang et al., 2021), provide evidence that EEG-based neural entrainment can be used to index the online process of word identification during SL. This measure provides valuable insights into the speech segmentation process, complementing the traditional offline learning outcome approaches.

1.3. *Individual differences in statistical learning*

Many SL studies report individual differences among participants, which can be quantified as either differences in learning outcomes, or differences in learning speed or trajectories (Bogaerts et al., 2022). This indicates that SL is not a capacity that everyone intrinsically possesses to the same degree or that follows the same timeline of learning (e.g., Batterink & Paller, 2017; Erickson & Thiessen, 2015; François, Tillmann et al., 2012; Misyak & Christiansen, 2012; Misyak et al., 2010; Siegelman, 2020; Siegelman & Frost, 2015).

There are also indications that SL ability is associated with individual differences in language acquisition, particularly delays or disorders in language development (Evans et al., 2009; Gabay et al., 2015; Lammertink et al., 2017; Newman et al., 2016; Singh et al., 2012; Vandermosten et al., 2019; Zhang et al., 2021). Specifically, earlier research found a relationship between SL in speech segmentation experiments and vocabulary development in children (Evans et al., 2009; Newman et al., 2016; Singh et al., 2012). In these (longitudinal) experiments, SL performance correlated positively with vocabulary size. Moreover, several studies point to a SL deficit in individuals diagnosed with developmental language disorder (DLD; e.g., Evans et al., 2009; Lammertink et al., 2017). On the other hand, the evidence for a SL deficit in developmental dyslexia (henceforth 'dyslexia') is mixed, with some studies finding evidence in favor of a SL deficit or delay in dyslexia (Gabay et al., 2015; Kerkhoff et al., 2013; Vandermosten et al., 2019; Zhang et al., 2021) while other studies do not find a difference between dyslexia and control groups for SL (Schmalz et al., 2017; van Witteloostuijn et al., 2019). The available evidence in favor of SL abilities predicting vocabulary outcomes as well as deficits in language disordered populations have yielded theories of individual differences in SL as an important predictor of language acquisition, including in the typically developing population (e.g., Conway et al., 2010; Erickson & Thiessen, 2015; Misyak et al., 2010; Siegelman, 2020).

If SL is indeed an important predictor of language development, an open question is: what underlies individual differences in SL, which in turn might predict inter-individual variation in language attainment? In order to better understand how language learners solve the speech segmentation problem, and why some individuals do this with ease while others might struggle – which may even culminate into a language impairment – we need to know more about the *underpinnings* of individual differences in SL. We fundamentally map SL as a multifaceted construct involving multiple cognitive and task-related components that might predict the individual differences in SL (Arciuli, 2017; Bogaerts et al., 2022; Siegelman, 2020; Siegelman & Frost, 2015). This is not to argue that an individual's SL capacity can be explained entirely by other cognitive factors, but we commit to the idea that SL can be influenced by them in a multi-faceted and complex manner (following Erickson and Thiessen (2015), for instance). This influence can lead to either facilitation or impairment of the SL process and thus predict inter-individual variability on SL tasks. We now turn to the question of which cognitive components are plausible candidates to influence individual differences in SL.

1.4. *Cognitive abilities and statistical learning abilities*

Multiple cognitive abilities have been theorized to contribute to individual differences in SL. One such ability is working memory (Arciuli, 2017; Misyak & Christiansen, 2012; Smalle et al., 2022). However, in contrast to theoretical proposals, previous empirical research has not found conclusive evidence that individual differences in working memory predict domain-general SL ability. Studies either failed to find significant correlations at all (Conway et al., 2010; Siegelman & Frost, 2015), or found a relation only for SL of adjacent patterns but not for SL of non-adjacent patterns[4] (Misyak & Christiansen, 2012). Moreover, Smalle et al. (2022) used a different method that not only *measured* individuals' working memory capacity but *overloaded* it, and interestingly found a significant improvement of SL ability for implicit word segmentation when high cognitive demand was induced. In contrast, Palmer and Mattys (2016) also imposed a cognitive load task on their participants, and found disrupted SL.

Another individual ability that has more recently been associated with speech segmentation is audio-motor synchronization. Assaneo et al. (2019) demonstrated that SL is better in individuals who show enhanced synchronization to an auditory speech rhythm on a behavioral level compared to individuals who do not synchronize. They developed a new task called the Speech-to-Speech Synchronization (SSS) task (further details of the task protocol: Lizcano-Cortés et al., 2022), where participants are instructed to repeat a

[4] Adjacent patterns are transitional probabilities between neighboring items such as syllables used for word segmentation, thus the probability of *XY* given the overall frequency of *X* (previously explained in section 1.1). Non-adjacent dependencies have intervening items, consisting of patterns like *X*[*Z*]*Y*, where *X* predicts *Y* over intervening *Z*.

whispered *'tah'* while listening to an isochronous[5] randomized stream of syllables and recall if certain syllables were presented in the stream. Crucially, participants are not explicitly instructed to synchronize their whispering to the rhythm of the syllable stream, but it turns out that some do. This task revealed a bimodal distribution of individuals, where participants could be divided into high and low synchronizers. High synchronizers – i.e., those who spontaneously adjusted their speech rhythm to the rhythm of the input – subsequently performed better than low synchronizers on a separate behavioral speech segmentation SL task. Furthermore, in a subsequent passive listening phase while recording *magnetoencephalography* (MEG), high synchronizers showed greater neural phase-locking to an external rhythmic syllable stream, specifically in the left inferior and middle frontal gyri, relative to low synchronizers. Additionally, differences in neural structure were found between groups, with the high synchrony group showing enhancement of the arcuate fasciculus white matter tract connecting the auditory and motor cortices. Moreover, the authors also found a significant correlation between white matter volume in the left arcuate fasciculus and the brain-to-stimulus synchronization. Thus, relative to low synchronizers, high synchronizing individuals, defined as those who spontaneously synchronize their speech rhythm to an external speech rhythm more closely: (1) showed greater neural phase-locking to the rhythm of spoken input during passive listening, (2) showed enhanced white matter connectivity between auditory and motor cortices, which significantly correlated with brain-to-stimulus synchronization, and (3) performed better in a behavioral SL word segmentation task. The authors hypothesized that the high synchronizers' increased neural entrainment reflects the synchronization of attentive processing to syllable onsets and facilitates speech parsing. This would then lead to better extraction of the transitional probabilities between syllables, underlying successful word segmentation.

Finally, another body of research indicates that musical training positively influences both speech and music processing, as well as SL (François, Chobert et al., 2012; Mandikal Vasuki et al., 2017; Schön & François, 2011; Shook et al., 2013). Specifically, François et al., 2012 conducted a two-year longitudinal study in which they compared effects of musical versus painting training on SL ability in two groups of 8-year-old children (starting age). All children were tested on their SL performance segmenting a sung artificial language[6] at the beginning of the study, after one year, and after two years. Before training SL ability did not differ between the groups, but after two years SL performance significantly improved in the music-training group only, and not in the painting group. Interestingly, in a different publication, François et al., 2012 hypothesized that musical training may improve SL through strengthening and/or more efficient reorganization of the auditory dorsal pathway. This dorsal pathway, originally proposed by Hickok and Poeppel (2007) as part of their dual-stream model of language processing, maps sensory (phonological) representations from the auditory cortex onto articulatory motor representations in the motor cortex. It is hypothesized to be critical for spoken language acquisition; auditory-motor coupling is essential for learning how to speak (Hickok & Poeppel, 2007; Rodríguez-Fornells et al., 2009) and has been hypothesized to be a neural substrate of speech segmentation through SL (Rodríguez-Fornells et al., 2009).

1.5. *Rhythmic ability and statistical learning*

Importantly, the brain areas described in Assaneo et al. (2019) where the concentration of white matter was greater and where more neural synchronization was found in the high synchronizing group (left lateralized arcuate fasciculus; left inferior and middle frontal gyri) correspond to the left dorsal pathway (Poeppel and Assaneo, 2020). This converges with the hypothesis by François et al., 2012 that the dorsal pathway might be improved in musically trained individuals and that this might benefit SL for speech segmentation. However, Assaneo et al. (2019) noted that musical experience alone did not explain their bimodally distributed results. As musical ability has been found to be heritable (Gingras et al., 2015), it may also be the case that the dorsal stream is organized more efficiently as part of the neurological substrate of innate musical ability. For instance, Zuk et al. (2022) found significant correlations between white matter pathway volumes in infancy and subsequent musical aptitude. Moreover, they found significant correlations between musical aptitude and language measures, as well as direct correlations between language skills and the white matter tracts that also correlated with musical aptitude. The authors found no significant correlations involving the arcuate fasciculus – which is part of the aforementioned auditory dorsal stream – but indicate that "this is likely due to the reduced overall number of reliable reconstructions in these temporal neural pathways in infancy, resulting in an insufficient sample size ($n \leq 17$)" (p. 6). Taken together, white matter structures in similar areas are important for both language and music abilities, and already in infancy individual differences in volume of at least some of these structures can predict musical and linguistic aptitude. More imaging research and larger sample sizes are warranted to further investigate this.

A critical component of musical ability that was frequently linked to language outcomes is rhythm perception ability (Ladányi et al., 2020; Langus et al., 2023; Nitin et al., 2023; Zuk et al., 2022). Rhythmic structure such as the hierarchical organization of meters,[7] is a shared feature of language and music (e.g., Asano, 2022; Poeppel & Assaneo, 2020). Recent research shows that both musical rhythm and linguistic rhythm are processed through synchronization of neural oscillations to hierarchically nested frequencies that are present in both language and music (Daikoku & Goswami, 2022; Fiveash et al., 2021; Giraud & Poeppel, 2012; Liberto et al., 2020; Menn et al., 2022; Peelle & Davis, 2012; Poeppel & Assaneo, 2020; Tierney & Kraus, 2015). Furthermore, *rhythmic*

[5] Happening at regular intervals. In this case, all syllables were 222 msec long, creating a constant syllable frequency of 4.5 Hz (see Assaneo et al., 2019, p. 7).

[6] All studies reported in this section did not use purely speech stimuli, but all used stimuli that are (combined with) tones or Morse codes. To our knowledge, no experiment has explicitly made a connection between musical ability and SL of speech.

[7] Regular patterns of strong and weak beats.

ability – the ability to accurately detect and (behaviorally) synchronize to an auditory pulse – has been found to predict language development (Bekius et al., 2016; Ladányi et al., 2020; Langus et al., 2023; Nitin et al., 2023; Zuk et al., 2022). In addition, several studies indicate that atypical rhythm sensitivity correlates with linguistic impairments (Boll-Avetisyan et al., 2020; Caccia & Lorusso, 2020; Fiveash et al., 2021; Flaugnacco et al., 2014; Huss et al., 2011; Kraus et al., 2014; Ladányi et al., 2020; Sallat & Jentschke, 2015).

Previous literature points out that more precise phase-locking of neural oscillations to an auditory input is hypothesized to reflect optimal processing – as the syllable onsets align with the phase of neural oscillations (e.g., Assaneo et al., 2019; Peelle & Davis, 2012; Poeppel & Assaneo, 2020). As earlier mentioned, neural entrainment can also be used to measure individual SL ability online (e.g., Batterink & Paller, 2017, 2019; Moreau et al., 2022; Pinto et al., 2022). Is an efficiency in phase-locking perhaps supported by rhythmic abilities relevant for both music and language processing, such as rhythmic motor synchronization and deducing metrical structures? Neurally, this could be indicated by a strengthened dorsal pathway between the auditory and motor cortices. Thus, is specifically *rhythmic ability* an underlying mechanism supporting SL, and are neural oscillations phase-locking to the rhythm of an auditory stimulus the neural mechanism indicative of SL during speech segmentation?

1.6. *Current study*

The aim of the current study is to contribute to the understanding of the neurocognitive underpinnings of individual differences in auditory SL for word segmentation. We investigated SL both online during familiarization by quantifying neural entrainment to the underlying statistical structure of the speech input, as well as offline in behavioral word recognition tasks in the test phase. Online measurement of SL was performed using EEG and the frequency-tagging methodology similar to earlier publications (e.g., Batterink & Paller, 2017, 2019; Moreau et al., 2022; Pinto et al., 2022). The current study is an extension of prior work in multiple ways. In order to investigate individual differences, we measured participants' performance on tasks assessing musical, rhythmic, linguistic, and general cognitive abilities. We then related these scores to the neural measure of SL. To our knowledge, a relation between musical/rhythmic abilities and SL specifically for word segmentation has not previously been researched. Furthermore, the online EEG entrainment measure of SL also has not yet been related to tasks assessing individual differences. See Fig. 1 and the paragraphs below for our predictions regarding the individual differences and SL.

We predicted that rhythmic abilities would positively correlate with SL performance. We tested rhythm perception using two tasks (Harrison & Müllensiefen, 2018a, 2018b; Zentner & Strauss, 2017). We predicted these tasks to be positively correlated, but we used multiple tasks to be sure that we measured rhythm perception as accurately as possible. We also measured behavioral rhythmic speech-to-speech entrainment by using the SSS task (Assaneo et al., 2019). We expected performance on this task to also be a predictor of SL, which would replicate a key finding reported by Assaneo et al. (2019). We further investigated interrelations between these rhythm tasks, the SSS task, and SL ability (see section 2.6 for details). In addition, we exploratively added a questionnaire about general musical ability and musical training experience (Bouwer et al., 2016; Müllensiefen et al., 2014).

Moreover, we broadened our search for individual differences in SL to general cognitive abilities by adding the forward Digit Span (Wechsler, 2008) as an indication of working memory capacity. We chose to use the forward Digit Span and not the backward Digit Span because the forward span is associated with verbal working memory and depends on the phonological loop, which is the most relevant for our study. The backward Digit Span, however, is more associated with executive functioning and cognitive control (e.g., Ostrosky-Solís & Lozano, 2006). As earlier studies mentioned in 1.4 did not find conclusive evidence on a connection between working memory and SL using post-learning tests, we exploratively investigated whether working memory aids SL online.

In addition, we administered a vocabulary test (Dunn & Dunn, 1998; Schlichting, 2005), adding to the earlier mentioned body of research with children (Evans et al., 2009; Newman et al., 2016; Singh et al., 2012) and extending this question into adulthood. Misyak and Christiansen (2012) have also assessed vocabulary in adults, where it correlated marginally with print exposure but not with SL. However, their vocabulary assessment differed from ours – described in 2.3.3.*d* – in that it required participants to choose a synonym for a target word, whereas our vocabulary test required participants to choose a picture corresponding to the meaning of a target word. Therefore, analogous to earlier research with children, we predicted a positive relation between SL and vocabulary size.

Finally, even though this experiment aimed to answer the new questions above, it is also a partial replication and extension of earlier experiments (Assaneo et al., 2019; Batterink & Paller, 2017; Pinto et al., 2022). We therefore expected to find comparable results to these earlier studies, consisting of increasing phase-locking to the word-frequency over the course of exposure in the structured condition, but not in an unstructured random condition (Batterink & Paller, 2017; Pinto et al., 2022). We also predicted a replication of the behavioral results of Batterink and Paller (2017) in the tasks of explicit and implicit memory of the words, which would also be in line with our pilot results (appendix B in the supplementary data). Moreover, we tested if the neural measure of SL correlated positively with the behavioral tasks (Batterink & Paller, 2017). We extend this prior work because the participants in the current study were speakers of Dutch, and the stimuli we used were newly created and adhere to Dutch phonotactics.[8] Finally, we expected to replicate the finding of an SL advantage in participants with a higher synchronizing ability as expressed by the *phase-locking value* (PLV) of their speech in the SSS task (Assaneo et al., 2019).

[8] More details on the methodology used to create these stimuli are described in van der Wulp et al. (2022). See also appendix B in the supplementary data for details on a pilot experiment with these stimuli.

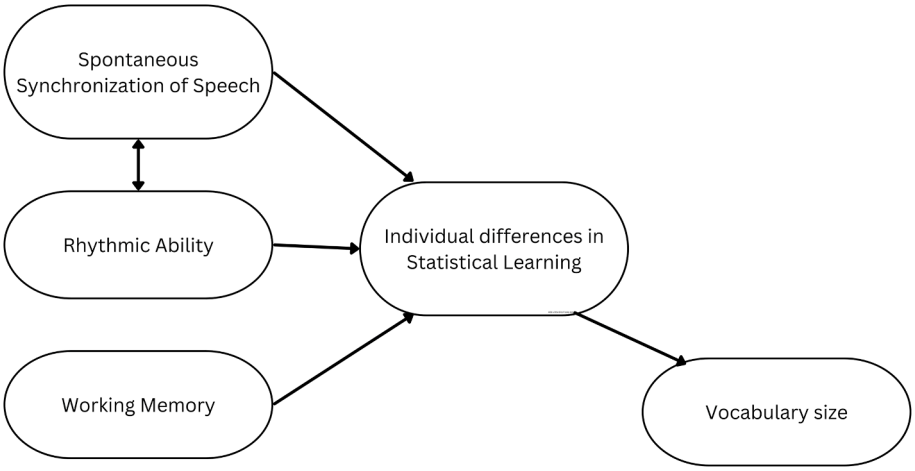

Fig. 1 – Predictions of the current study represented graphically.

2. Materials and methods

2.1. *Participants*

A total of 106 adults (88 F; 17 M; 1 X) participated in this study. Due to an unexpected termination of the test session, one participant only completed the Gold-MSI and the SL-part of the experiment and thus has missing data for the other tasks for individual differences. Participants were all native speakers of Dutch and between 18 and 35 years old ($M = 23.37$, $SD = 4.22$). Participants attended university ($N = 101$) or applied university ($N = 5$) as their highest educational level.

The experiment was approved by the Linguistics Chamber of the Faculty Ethics Assessment Committee of Humanities at Utrecht University (reference number: LK-22-174-02), and participants were compensated with a €20 gift card for their time (the session took approximately two hours).

2.1.1. *Bayesian updating procedure*

We started with an initial sample of 45 participants, identical to Batterink and Paller (2017). Then, we performed Bayesian Updating (Rouder, 2014), by repeating the statistical analyses after every added sample of 15 participants, until the threshold value of a Bayes Factor (BF_{10}; Jeffreys, 1961) > 6 or < 1/6 would be reached for our *critical analyses*, or when we would reach a maximum sample of 105 participants.[9] The latter was the case. The critical analyses (marked green in the study design table in appendix A in the supplementary data) were the following:

- The analysis for the replication of the EEG results of Batterink and Paller (2017; see section 2.4.1), with regard to a difference in the WLI between the structured and random conditions. We already found a $BF_{10} > 1000$ in the first sample of $N = 45$, which stayed that large with every update (see RStudio supplement).
- The correlations between the tests for rhythmic ability (see section 2.6), in order to be able to perform the mediation analysis. This is the analysis that increased the sample to $N = 105$. One of these correlations did not yield a $BF_{10} > 6$ until our final update (see section 3.3.).
- Evidence for or against a direct effect of audio-motor synchronization (Assaneo et al., 2019) on the $WLI_{structured}$, in order to be able to perform the mediation analysis (see section 2.6). We did not perform this analysis until our final sample (see section 3.4.).

2.1.2. *Exclusion criteria*

Participants were not invited to participate if they reported having a history of hearing impairments or tinnitus, AD(H)D, other attention or concentration issues, dyslexia, or other language-related impairments. Furthermore, data of participants was excluded for a certain task after participation in the case of technical issues, which was the case for some participants ($N = 3$) with the SSS task, where the stimuli were audible in the recording and masked the participants' whispers. This made the PLV calculation impossible for that task. Furthermore, data from one participant was excluded for the target detection task because they detected fewer than 50% of targets (26.30% detected).

2.2. *Stimuli*

The stimuli consisted of syllables which were combined into tri-syllabic nonwords (from now on referred to as 'words') that adhered to Dutch phonotactics and have been piloted for their learnability (see appendix B in the supplementary data for details on the pilot experiment). The syllable inventory consisted of 12 syllables, from which four words were formed for the *structured condition*:/suχita, tobamø, sytøbo, χøbyti/. In the structured stream, the transitional probabilities of neighboring syllables were 1.0 within a word and .33 between

[9] Due to the one participant with missing data for the individual differences tasks except the Gold-MSI, we collected data from 106 instead of 105 participants. See Stage 1 for simulation-based estimates of statistical power: https://osf.io/2y6sx.

words. The word order was pseudorandomized, such that the same word did not repeat consecutively. More details on the methodology used to create these stimuli are described in van der Wulp et al. (2022).

We also created a corresponding random stream (Batterink & Paller, 2017), which forms the *random condition*. In the random condition, a different set of 12 syllables was concatenated in a pseudorandom order, under the constraint that the same syllable could not consecutively repeat (as in Batterink & Paller, 2017). This yielded a transitional probability of .09 throughout the random condition. The syllables used in this condition were:/da, pø, nu, dø, χo, py, ro, dy, sa, χy, ri, sø/, corresponding to set *B* in the pilot experiment (see supplementary data: table C1, and see van der Wulp et al. (2022) for more details on the methodology used to create these stimuli).

The stimulus lists were converted to concatenated speech without pauses using MBROLA diphone synthesis (male Dutch voice nl2, at a monotone F0 of 100 Hz; Dutoit et al., 1996). All syllables were 300 msec long (100 msec consonant, 200 msec vowel), creating a word-length of 900 msec. Thus, this yielded a syllable frequency of 3.3 Hz and a word or triplet frequency of 1.1 Hz (see Fig. 2). We generated coarticulated speech streams of 13.5 min per condition in total, divided over three blocks of 4.5 min. Each block was made up of 900 syllables (300 words).

We used GoldWave (GoldWave Inc., 2022) to add a linear fade-in and fade-out of 1.5 s at the beginning and end of each block, to avoid a segmentation cue at the beginning of the stream. Stimuli were presented with Presentation (www.neurobs.com). Finally, we used GoldWave to add a cue point[10] at the onset of each syllable in the continuous audio files, so that they could be read as EEG markers with Presentation. The EEG markers and their corresponding syllables can be found in table C1 in appendix C in the supplementary data.

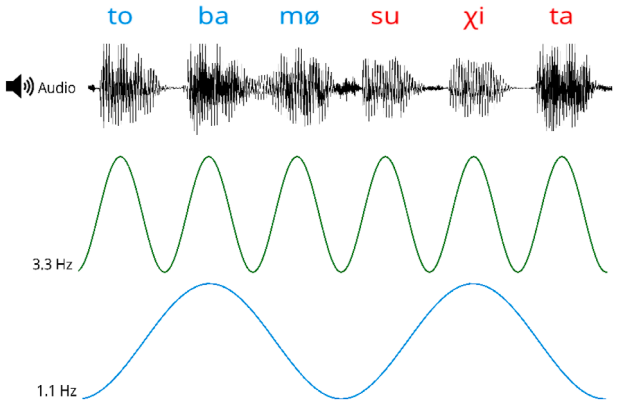

Fig. 2 – Stimuli and stimulus frequencies in the structured stream. The audio represents the depicted syllables. The syllables of the same color form a word. The green waveform depicts the syllable frequency of 3.3 Hz. The blue waveform depicts the tri-syllabic word frequency of 1.1 Hz.

[10] For more information about cue points, see this manual.

2.3. *Procedure*

A schematic depiction of the experimental procedure can be viewed in Fig. 3. Detailed descriptions of the procedure are given in the following sections.

2.3.1. *Listening task*

Participants first performed the listening task in the structured condition. After this, the rating task and target detection task (TDT; see 2.3.2.) were administered, followed by another iteration of the listening task to the random stream. The

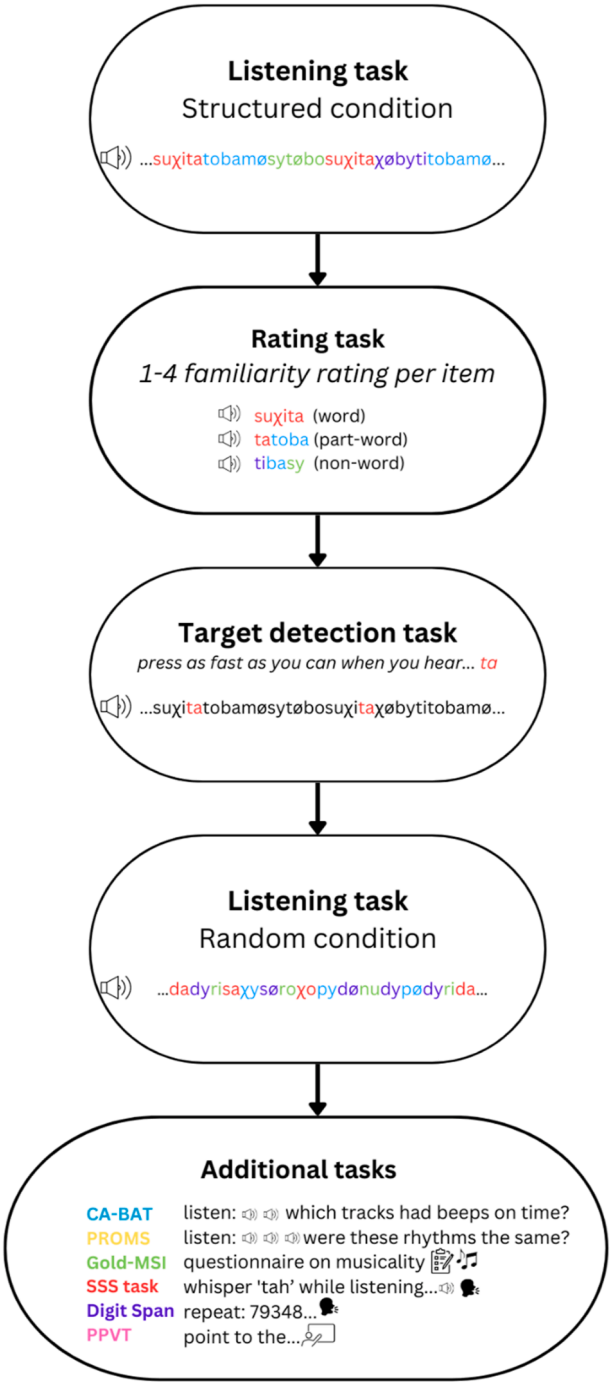

Fig. 3 – Schematic overview of the experimental procedure.

listening task was divided into three blocks of 4.5 min per condition, yielding 13.5 min per condition and 27 min in total for both conditions. Participants took short self-timed breaks between blocks.

2.3.2. *Behavioral tasks of SL outcomes*

Following the structured condition of the listening task, participants performed two tasks to assess their explicit and implicit knowledge of the words: a familiarity rating task and a reaction-time based target detection task (TDT).

With respect to the rating task, participants were auditorily presented with a word or foil in each trial. The foils could be of two kinds: one being a *part-word* spanning a word boundary from the stream, or a *non-word* made up of syllables from the stream but recombined in an order that never appeared (see Fig. 3; see table C2 in appendix C in the supplementary data for the full list of foils). There were 16 trials consisting of the four words from the listening task, all eight possible part-words and four non-words. On each trial, participants rated on a four-point scale how familiar the word was to them (scale: unfamiliar – fairly unfamiliar – fairly familiar – familiar).

The second post-learning task our participants performed was the TDT (Batterink, 2017; Batterink et al., 2015; Batterink & Paller, 2017, 2019). Participants were presented (auditorily and visually) with a target syllable and subsequently heard a shortened version of the structured condition from the listening task, containing 16 words (4 words each repeated 4 times) corresponding to 48 syllables, and the same word not repeated in succession. They were asked to press a button as quickly and accurately as possible when they heard the target syllable. For each target syllable there were three speech streams, with the target occurring four times per stream, resulting in 36 speech streams and 144 targets for this task.

2.3.3. *Additional tasks for individual differences*

a. Musical and rhythmic abilities

We employed three measures assessing rhythmic and musical abilities of the participants. First, participants performed the Computerized Adaptive Beat Alignment Test (CA-BAT; Harrison & Müllensiefen, 2018a; 2018b), in which participants listened to the same piece of music twice, accompanied by beeps in two conditions. In one condition, the beeps were synchronized with the rhythm of the music, and in the other condition, the beeps were not synchronized with the rhythm of the music. Participants indicated which of the two tracks had the beeps in sync with the rhythm of the music.

Second, participants completed the Rhythm and Accent sub-tests of the short version of the Profile of Music Perception Skills (PROMS; Zentner & Strauss, 2017). In this task, participants listened twice to the same rhythm and then to a third rhythm. Participants then indicated whether the third rhythm was identical or different compared to the first two.

Third, participants completed a self-report questionnaire of general musical abilities: the Goldsmiths Musical Sophistication Index (Gold-MSI; Müllensiefen et al., 2014), translated to Dutch (Bouwer et al., 2016). The questionnaire consisted of the following sub-scales: active engagement with music, perceptual abilities, musical training, singing abilities and emotional engagement. Participants filled out this questionnaire during EEG set-up.

b. Spontaneous Synchronization to Speech

We administered the implicit fixed version of the Speech-to-Speech Synchronization (SSS) task (Assaneo et al., 2019; Lizcano-Cortés et al., 2022), in which participants were instructed to whisper '*tah*' while listening to an isochronous stream of syllables and recalling which syllables were presented afterwards. We had translated the instructions to Dutch for our sample of Dutch native speakers.

c. Working memory

Participants performed a forward Digit Span (Wechsler, 2008) as an indication of working memory capacity. In this test, the experimenter orally named digits and the participant was instructed to repeat them. The number of digits increased until the participant failed to remember two digit-series of the same length.

d. Vocabulary

Finally, we administered the Dutch Peabody Picture Vocabulary Test, third edition (PPVT–III–NL; Dunn & Dunn, 1998; Schlichting, 2005) to measure the vocabulary size of our participants. The PPVT–III–NL is a task where participants are presented with a word and four pictures. The participant then indicates which picture corresponds to the meaning of the word. The test is suitable for ages 2; 3 through 90 years and is norm-referenced for both the infant and adult population.

2.4. *EEG recording and analyses*

During the listening task, EEG was recorded at a sampling rate of 512 Hz using 64 Ag/AgCl-tipped electrodes attached to an electrode headcap 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, above and below the left eye, and at the outer canthi of both eyes. Scalp signals were recorded relative to the Common Mode Sense (CMS) active electrode and then re-referenced during data analysis to the average of the mastoid electrodes. Impedance of the channels was kept below 20 mV. If the impedance of a channel was higher than this, it was labeled as a bad channel during data collection to be interpolated during data analysis.

The EEG data was analyzed in MATLAB (The MathWorks Inc., 2019) using EEGLAB (Delorme & Makeig, 2004) and the ERPLAB open-source toolbox (Lopez-Calderon & Luck, 2014). The data was bandpass filtered from .1 to 30 Hz[11] and 50 Hz notch filtered offline. Bad channels identified upon visual

[11] Sixteen participants had slow drifts in their data. This made the manual artifact rejection difficult. Therefore, their data was filtered from .5 to 30 Hz instead. This did not influence their ITC results at the frequencies of interest.

inspection of the data or during data collection were interpolated (mean *N* of interpolated channels structured = 5.18; random = 6.49). Data sections comprising large artifacts were also identified through visual inspection and manually rejected using the EEGLAB plugin VisEd (Desjardins et al., 2019). A channel was labeled as bad during the analysis if it was labeled bad during data collection due to high impedance, or if it showed frequent noise or drifts upon visual inspection of the data. Eye movement artifacts were retained, as they are not time-locked to the stimulus onsets and have a broad power spectrum that does not affect the narrow-band neural oscillations (Srinivasan & Petrovic, 2006).

We time-locked the data to the onsets of the tri-syllabic words and divided it into non-overlapping epochs of 10.8 sec, corresponding to the duration of 12 trisyllabic words (36 syllables). We then quantified phase-locking to the word (1.1 Hz) and syllable (3.3 Hz) frequencies using the ITC, which ranges from 0 to 1. An ITC of 1 indicates perfect phase-locked neural activity to a given frequency, and 0 indicates no phase-locking at all to that frequency. The ITC was calculated after a Fast Fourier Transform (FFT) for each epoch across frequency bins of interest: between .6 and 5 Hz, with a bin width of .09 Hz (following Batterink & Choi, 2021; Benjamin et al., 2021; Moreau et al., 2022). The Word Learning Index (WLI) was then calculated as a mean for each participant over the entire exposure period, as well as for each epoch bundle over the time course of exposure, for both the structured and random conditions.

$$WLI = \frac{ITC_{word\ frequency}}{ITC_{syllable\ frequency}}$$

To perform the time course analysis, we followed the methodology of Moreau et al. (2022) using a sliding window to map learning trajectories during the listening task. We created *epoch bundles* each containing 5 epochs, with each bundle shifted by one epoch (e.g., epochs 1–5, 2–6, 3–7, etc.). This resulted in 54 sec of exposure per bundle. We computed the ITCs and WLI for the 20 fronto-central electrodes previously used by Moreau et al. (2022).[12]

2.4.1. Statistical analyses of the neural data

We statistically tested for evidence for the alternative hypothesis (H1) by calculating the Bayes Factor (BF), adhering to an inference threshold of $BF_{10} > 6$. Correspondingly, inference of evidence for the null hypothesis (H0) is expressed as $BF_{10} < 1/6$. However, the BF is continuous, and can be interpreted as such. The higher the BF is, the more evidence we have for H1, and the smaller the BF, the more evidence for H0 (see also Dienes, 2019; Schmalz et al., 2023). We calculated the ITC for the word and syllable frequencies over the exposure period and used them to compute the WLI, as described in 2.4 above. We then conducted our statistical analyses using R (R Core Team, 2021) and by creating Linear Mixed Models (LMM) with the packages tidyverse (Wickham et al., 2019), lme4 (Bates et al., 2015), and lmerTest (Kuznetsov et al., 2017). The model for the neural data had the WLI (centered around 0 by subtracting the mean) as the dependent variable and we initially included a random slope for language condition (structured/random) per participant. We expected the WLI to be higher in the structured than in the random condition, and to increase as a function of exposure during the listening task in the structured but not in the random condition, replicating earlier findings (Batterink & Paller, 2017; Moreau et al., 2022; Pinto et al., 2022; Van der Wulp, 2021). We statistically determined this by including condition as a predicting factor and subsequently an interaction of condition and epoch bundle.[13]

We then computed two Bayes Factors – one for the main effect of condition and one for the interaction – using Dienes (2008) calculator method (implemented in R by Baguley & Kaye, 2010). In the calculator, H0 is modelled as a point estimate (i.e., 0 is the only plausible value) and H1 is modelled as a distribution representing the probability of different magnitudes of the effect if H1 is true. Specifically, we used a half-normal distribution with the mode set to 0 and the standard deviation set to *x* where *x* is an estimation of the predicted effect. For the effect of condition, we set *x* = .19, as was the estimate of this effect in Batterink and Paller (2017; see re-analysis in Stage 1 code supplement). For the interaction, we set *x* = .01, which is the estimate of the interaction with epoch bundle in Moreau et al. (2022, Table S3).[14]

For each Bayes Factor test, the Dienes calculator needs two numbers which provide a summary of the data, specifically, a mean and a standard error. We followed Silvey et al. (2024) and used the β and SE of the relevant coefficients (i.e., for the main effect of condition and the interaction) extracted from the mixed effects model. See the simulation supplement from Stage 1 for the models yielding these estimates on the data of Batterink and Paller (2017). If we encountered singularity errors or if the model did not converge, we first removed the correlations between random slopes. If it still did not converge or still was singular, we removed the random slope.

We followed the analyses with sensitivity analyses by reporting Robustness Regions (Dienes, 2019). Robustness Regions provide the range of predicted values we could have set as *x* (i.e., the SD of the model of H1), while still drawing the same qualitative conclusion with respect to our data. So, for example, if we obtain a $BF_{10} > 6$, and thus conclude there is robust evidence for H1, what range of values of *x* could we have used and have obtained a BF at least as large as 3 (indicating moderate evidence)? Or if we obtain a $BF_{10} < 1/6$, what range of values of *x* we could have used and found a $BF_{10} = 1/3$ or less? When computing the range, we considered only

[12] F3, F1, Fz, F2, F4, FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2 & CP4.

[13] Our manual artifact rejection method yielded variations in the *N* of epoch bundles per participant and condition. Moreover, the data could be rejected at any moment in the EEG file i.e., epoch bundle 10 for participant A could be at a different timepoint than epoch bundle 10 for participant B. Therefore, we used the first syllable number in the bundle ('ur-event') as the predictor in the models examining entrainment over time.

[14] The prior for this analysis preregistered at Stage 1 was x = .07, which was based on the block β from Batterink and Paller (2017). However, if we hypothetically had a similar increase over time in both datasets, the β for block would be larger because it occurs over a longer period of time, whereas an epoch bundle represents a much smaller increment of time. We also exploratively repeated the analysis by dividing the data into the three exposure blocks and running the model on the WLI per block under this block prior (see S.1 in the Supplementary Materials).

plausible values and therefore we looked at values between 0 and .38 and between 0 and .02 for the condition and interaction effects, respectively. These values are in each case twice as large as the effects found by Batterink and Paller (2017) and Moreau et al. (2022). In theory the WLI can range to infinity, but we did not expect the effect to be more than twice as large as in previous studies.

2.5. Behavioral data analyses

2.5.1. Group analyses of behavioral SL outcome measures

The dependent variable for the rating task consisted of the familiarity ratings on the four-point scale. Random effects were random intercepts for participant and item. We tested whether words were judged as more familiar than part-words and subsequently non-words by using a Cumulative Link Mixed Model (CLMM) from the R package *ordinal* (Christensen, 2022) with familiarity rating as the dependent variable and word category as predictor. Because the rating task has not been analyzed with a CLMM before, we used the package Bain, which stands for *BAyesian INformative hypothesis evaluation* (Gu et al., 2021; Hoijtink et al., 2019). Bain computes the approximate adjusted fractional BF. According to Gu et al. (2014) and further elaborated in Gu et al. (2018) the prior distribution of the structural parameters can be chosen as:

$$h(\boldsymbol{\theta}) = N\left(\mathbf{0}, \sum\nolimits_{\infty}\right) \tag{1}$$

where, $\boldsymbol{\theta}$ contains the parameters that are evaluated in the hypothesis that is presented below, $\mathbf{0} = (0, \ldots, 0)^{T}$, and $\sum_{\infty}$ equals $\sum_{\boldsymbol{\theta}}$ (see below) rescaled such that the variance of each parameter is approaching infinite, such that the impact of this prior distribution on the posterior is negligible as the posterior only depends on the data. Subsequently, the posterior distribution is approximated by a normal distribution:

$$g(\boldsymbol{\theta}|\mathbf{X}) \approx N\left(\hat{\boldsymbol{\theta}}, \sum\nolimits_{\boldsymbol{\theta}}\right) \tag{2}$$

Where $\mathbf{X}$ denotes the data, $\hat{\boldsymbol{\theta}}$ denotes the estimates of structural parameters, and $\sum_{\boldsymbol{\theta}}$ denotes their covariance matrix (Gu et al., 2014, p. 516). Finally, the BF is represented for a given hypothesis H_i against an its complement *Hc* as the ratio of the posterior and prior probabilities that the inequality constraints hold:

$$BF_{ic} = \frac{f_i}{c_i} \times \frac{1 - c_i}{1 - f_i} \tag{3}$$

where c_i called complexity is the proportion of the prior distribution (Equation (1)) in agreement with H_i, and f_i called fit is the proportion of the posterior distribution (Equation (2)) in agreement with H_i (Gu et al., 2014, 2018). Note that, H_c is the complement of H_i, that is, "not H_i." By taking the foils as intercept, we formulated the following informative hypothesis for Bain, which was evaluated against its complement (Equation (3)):

H1. $\beta_{part\text{-}word} > 0$ & $\beta_{word} > 0$ & $\beta_{word} > \beta_{part\text{-}word}$.

After the initial analysis, we also conducted a sensitivity analysis. In Bain, this is done by increasing the size of the fraction *b* of information in the data used to specify the prior variance from $1 \times b$ (default), to $2 \times b$, as well as $3 \times b$. If the BF does not substantially change, we can conclude that the results are robust (Hoijtink et al., 2019, pp. 548–549).

With respect to the TDT, RTs were only taken into consideration for any of the analyses if the button press occurred within 1200 msec after the target onset, as has been done in previous studies (Batterink, 2017; Batterink & Paller, 2017, 2019). All other responses are considered false alarms. Reaction times (RTs) were analyzed using a LMM with RT as the dependent variable and within-word syllable position (word-initial, word-medial, and word-final) as the predicting factor, to establish if the facilitating effect towards the word-final syllable is present in our data. We furthermore added a random intercept for participant to account for individual differences in baseline RTs. Finally, we added the variable syllable repetition as a covariate, referring to the trial number of the target syllable in the stream (1–4),[15] in order to control for an increase in RTs over the course of the stream that has been observed previously (Batterink, 2017; Wang et al., 2023). We used the same methodology for calculating the BF as in 2.4.1, with our model of H1 as a half-normal distribution with a mean of 0 and an SD of 31.91, which was the result of our pilot experiment on the TDT (see appendix B in the supplementary data).

We followed this analysis with a sensitivity analysis reporting a robustness region (Dienes, 2019). We tested for prior models of H1 where the RT difference is 0–150 msec to find the region where the BF_{10} is still > 3 or < 1/3. In our pilot, we observed an effect of 31.91 msec, thus this maximum is large in comparison. However, a difference of 150 msec is theoretically plausible, as the fastest RT for the third syllable in our pilot was around 400 msec and an average button press takes about 250 msec. Thus, 400–250 = 150 msec is the maximum effect we could theoretically expect.

2.5.2. Correlations between neural and behavioral SL data

For the rating task, we computed a composite *rating score* for each participant, following Moreau et al. (2022; Batterink & Paller, 2019), subtracting the mean rating for foils (part-words and non-words) from the mean rating score for words. For the TDT, we calculated a *RT facilitation score* for each participant (Batterink & Paller, 2019; Moreau et al., 2022), by subtracting the mean RTs for the third syllable from the mean RTs for the first syllable and dividing this by the mean RTs for the first syllable: ($RT\ facilitation = (RT_{S1} - RT_{S3})/RT_{S1}$), which accounts for individual baseline RTs. We conducted Bayesian correlation analyses between the overall WLI in the structured condition, the rating score, and the RT facilitation score to

[15] At Stage 1, we conceptualized this as syllable position; 1–48 since there are 48 syllables in each stream. However, we only placed cue points in the streams at target syllables, and therefore only these appeared in the log files. There were three streams per target (section 2.3.2.), and randomized which stream was presented when. Unfortunately, this information did not appear in the log files. Therefore, we only know if it was the 1–4th time a target was presented within one stream. This is conceptually similar to our initial plan, so we included this (1–4th presentation) as the covariate instead.

determine whether individual variability in neural entrainment during exposure is related to subsequent SL performance. We performed these correlations using the statistical software JASP (JASP Team, 2023). The prior distribution for correlations in JASP is described by a beta-distribution centered around zero and with a width parameter (κ) of 1 as the default (see Fig. 4). The width is inversely related to the parameters of the beta distribution. For instance, a prior weight of .5 generates a beta(2,2) stretched from −1 to 1 (2 = 1/.5). In this case, the beta distribution is truncated at 0, because we only hypothesized positive correlations. Since the effects in Batterink and Paller (2017) were $r = .32$ for the rating task, and $r = .42$ for the TDT, we adhered to the prior $\kappa = .5$, which places less prior weight on big effect sizes and relatively more around 0. We followed this analysis with a sensitivity analysis. In JASP, this feature is implemented, and the output shows the results for every possible value of κ (between 0 and 2). We interpret a result as robust when the BF_{10} does not drop below 3 when the prior varies.

2.5.3. *Analyses of behavioral tasks for individual differences*

The CA-BAT (Harrison & Müllensiefen, 2018a; 2018b) generates a score per participant according to the Item Response Theory. Essentially corresponding to z-scores, a score of 0 corresponds to the mean of the calibration sample and a score of 1 to the standard deviation of the calibration sample's rhythm discrimination ability.

The PROMS (Zentner & Strauss, 2017) yields a raw score for the rhythm subtest (between 0 and 8) and the accent sub-test (between 0 and 10), the mean of which we recorded as one data point per participant.

Self-reported musical experience and expertise as measured with the Gold-MSI questionnaire (Bouwer et al., 2016; Müllensiefen et al., 2014) yields a general score between 1 and 7 for each participant and sub-scores also ranging between 1 and 7 per sub-scale.

For the SSS task (Assaneo et al., 2019), we adhered to the protocol described in Lizcano-Cortés et al. (2022). We calculated the PLV for each participant's whispers to the input rhythm of 4 Hz.

With respect to the forward Digit Span test (Wechsler, 2008), we measured the longest span for each participant. This test yields a score between 1 and 16, which was recorded as one data point per participant.

Finally, for the PPVT–III–NL (Dunn & Dunn, 1998; Schlichting, 2005), scores were also recorded as one data point per participant. This score is the age-corrected WBQ (WoordBegripsQuotiënt – 'Word Understanding Quotient'), which is a quotient measure similar to intelligence (IQ). The calibrated mean vocabulary score per age is 100, and scores below 100 indicate less-than average performance, while scores above 100 indicate above average vocabulary size for the participant's age.

All scores on the individual differences' tests were standardized before statistical analyses were conducted. This was done by subtracting the mean from the variable, and subsequently dividing that by the standard deviation of the variable.

2.6. *Analyses of individual differences in statistical learning*

For the analyses of individual differences, we first computed correlations between all of our tests for individual differences: the CA-BAT, PROMS, SSS task PLV, Gold-MSI, Digit Span, and PPVT–III–NL. We performed these correlations using the statistical software JASP (JASP Team, 2023). With regard to the

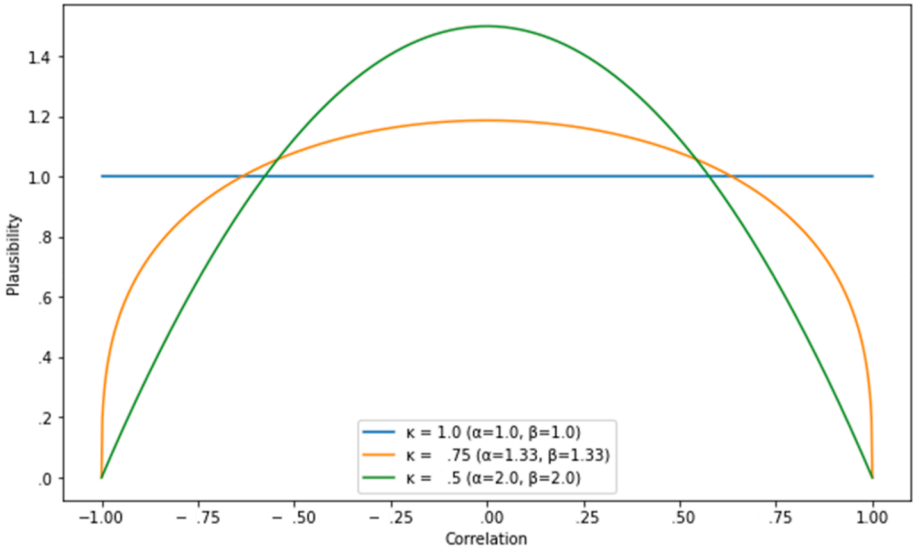

Fig. 4 – Beta prior distributions in JASP for correlations. In JASP, one specifies the width of the prior distribution (κ). The width is inversely related to the parameters of the beta distribution. The default value of κ is 1 (blue line). We used $\kappa = .5$ (green line) for medium and $\kappa = .75$ (orange line) for large, hypothesized correlations. When testing one-sided, the distribution is truncated at 0.

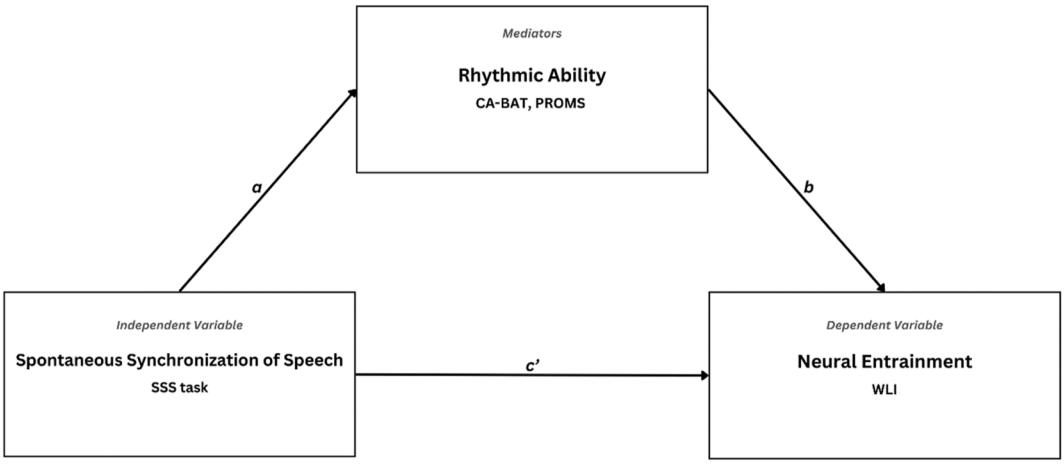

Fig. 5 – Mediation analysis planned at Stage 1, hypothesizing a direct effect of SSS PLV (spontaneous synchronization of speech) on the WLI (neural measure of SL) in the structured condition, adding the CA-BAT, PROMS (both rhythmic ability), as mediators. The c' path denotes the direct effect, and the path ab denotes the mediated effect. This analysis was not performed, as there was no direct effect (c' path, see section 3.4).

priors for these correlations, we expected the measures of rhythm (e.g., CA-BAT, PROMS, and SSS task PLV) to be highly positively correlated. Therefore, we used the prior $\kappa = .75$, which places relative weight on larger effect sizes. For more information on the prior distribution in JASP, see section 2.5.2. Exploratively, the Gold-MSI measuring general musicality was also hypothesized to show a positive correlation with the rhythm tasks, but we did not necessarily expect correlations between the Digit Span, PPVT–III–NL, and rhythm tasks. For these explorative correlations, we adhered to the prior $\kappa = .5$, which places less prior weight on big effect sizes and relatively more around 0. This gave us a reasonable chance of finding a theoretically interesting medium-to-large effect size (see also appendix A in the supplementary data). We followed these analyses with sensitivity analyses provided by JASP (see section 2.5.2).

Subsequently, in order to assess the influence of our predictors for individual differences on SL, we planned to perform a mediation analysis with multiple mediators (e.g., Dienes, 2019; Field, 2013; Zhang & Wang, 2017). The WLI in the structured condition was the dependent variable, and we predicted a direct effect of the SSS PLV based on earlier research (Assaneo et al., 2019). This would indicate that individuals with a higher PLV on the SSS task show more phase-locking to our frequencies of interest and also better SL. We tested for this direct effect initially by performing a regression of the SSS task on the $WLI_{structured}$, and subsequently loading the model in the package Bain (Gu et al., 2021; Hoijtink et al., 2019), under the informative hypothesis for the direct effect: *c-path* > 0. The hypothesis for a null effect was defined as *c-path* = 0. For an explanation of how Bain calculates the prior and posterior distributions, and the BF, we refer the reader back to section 2.5.1. We hypothesized that the direct effect, if found, would be mediated by one or more of our measures of rhythmic ability (see Fig. 5). Tasks that did not correlate with the SSS task, would be correlated separately with the $WLI_{structured}$ under the prior $\kappa = .5$, with sensitivity analyses as described in section 2.5.2. Since there was no direct effect (see section 3.4.), we did not perform the mediation analysis and instead correlated all tasks measuring individual differences with the $WLI_{structured}$ in this way.

3. Results

3.1. *EEG results*

We first calculated the ITC and WLI over the entire exposure period in each condition. We then plotted the ITC for the frequencies under 5 Hz (see Fig. 6). This yielded clear peaks at the syllable frequency in both structured and random conditions, and a peak at the word frequency in the structured condition only. The WLI was skewed ($W = .80$, $p < .001$), so we log-transformed the WLI before mean-centering.[16] We included a random intercept for participant, as the model did not converge with random slopes. We found extreme evidence for an effect of condition on the overall WLI ($\beta_{condition\ structured} = .17$, $SE = .03$, $BF_{10\ (0,\ .19)} > 1000$). We then computed the robustness region (see section 2.4.1.), which indicated a robust effect for the entire preregistered range $RR_{BF\ >\ 3}$ [.01, .38].

The model for condition in interaction with epoch bundle included a random slope for condition per participant. We found evidence for H0 ($BF_{10} < 1/6$, see section 2.4.1.) on the interaction ($\beta_{condition*epoch\ bundle} = -5.14 \times 10^{-7}$, $SE = 6.07 \times 10^{-6}$, $BF_{10\ (0,\ .01)} < .001$, $RR_{BF\ <\ 1/3}$[.001, .02]), indicating that the progression of the WLI across time did not differ by condition, in contrast to our original hypothesis. Fig. 7 shows the WLI as a function of exposure time per condition.

[16] The WLI in Batterink and Paller (2017) was also skewed, so we log-transformed it also when we constructed the prior at Stage 1. See the RStudio supplement for histograms of the distributions.

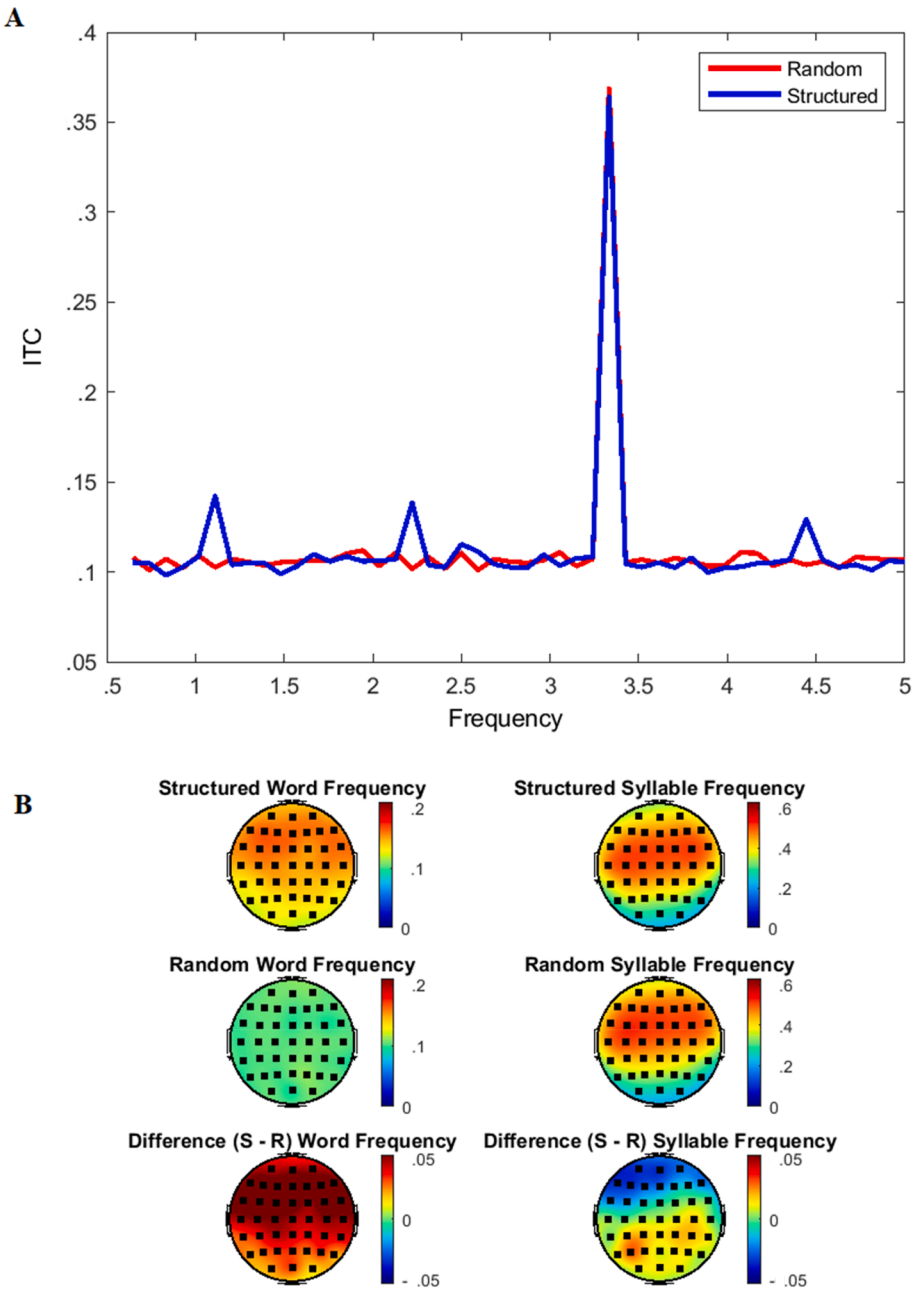

Fig. 6 – Descriptive results of the EEG analysis.
***Note.* A) Inter-Trial Coherence results per condition; B) Topographic distributions of the ITC per condition and frequency of interest. Note that different scales are used for word and syllable frequencies. Both A) and B) are calculated over the entire exposure duration.**

In order to further investigate why the WLI did not progress over time in the structured compared to the random condition, we separately plotted the time-courses of the ITC to the word and syllable frequencies (ITC_{word} and $ITC_{syllable}$; see Fig. 8). These plots indicate that the ITC_{word} does increase towards the end of the structured but not random condition. However,

the $ITC_{syllable}$ also fluctuates over time but does not decrease. This lack of decrease in $ITC_{syllable}$, which contrasts with the findings of Batterink and Paller (2017), has implications for the composite WLI (section 2.4). Therefore, we exploratively added a not-preregistered analysis where we tested whether the interaction between condition and epoch bundle was present for ITC_{word}. As the prior for this analysis, we took the estimate for this interaction for ITC_{word} in the adult group from Moreau et al. (2022, Table S3): 4.06×10^{-3}. We again used the methodology for calculating the BF as elucidated in 2.4.1. and calculated Robustness Regions between 0 and 8.12×10^{-3} (twice as large as the prior, in line with our other analyses). Results of this analysis again indicated evidence for H0 ($\beta_{condition*epoch\ bundle} = 9.95 \times 10^{-6}$, $SE = 3.69 \times 10^{-6}$,

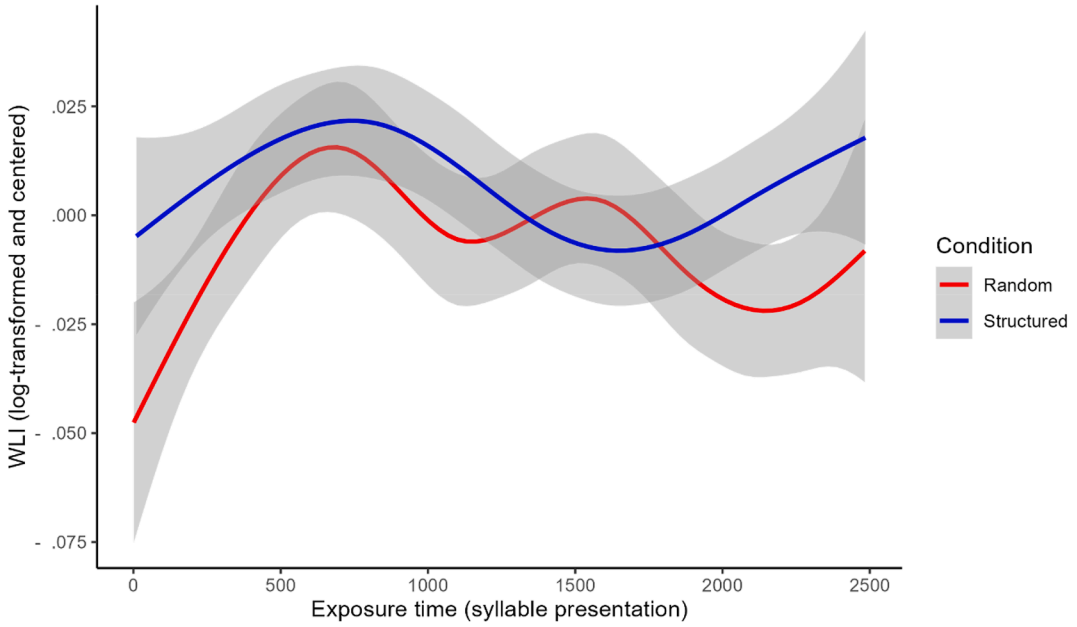

Fig. 7 – WLI over time per condition.

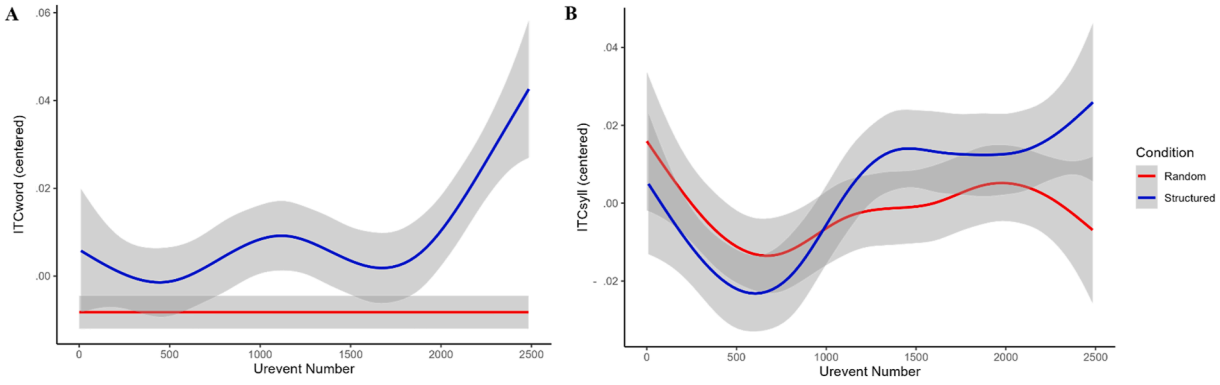

Fig. 8 – ITC_{word} and $ITC_{syllable}$ as a function of time per condition.

$BF_{10\ (0,\ 4.06\times10^{-3})} = .007$). The Robustness Regions clarified that the effect is smaller than the prior, yielding evidence for H0 $RR_{BF\ >\ 3}[8.3 \times 10^{-5}, 9 \times 10^{-3}]$.

Finally, we explored in a not-preregistered follow-up analysis whether ITC_{word} increased over time in the structured condition alone. We again based the prior on Moreau et al. (2022, p. 6 (Table 1): 2.14×10^{-3}) and calculated Robustness Regions between 0 and 4.28×10^{-3}. Results of this analysis indeed indicated evidence for an increase of ITC_{word} over time in the structured condition ($\beta_{epoch\ bundle} = 1.02 \times 10^{-5}$, $SE = 2.64 \times 10^{-6}$, $BF_{10\ (0,\ 2.14\times10^{-3})} = 6.96$, $RR_{BF\ >\ 3}[1.00 \times 10^{-5}, 2.73 \times 10^{-3}]$). Taken together, we did not find evidence for different learning trajectories for the preregistered dependent composite variable WLI. On the other hand, we did find an increase in the not-preregistered dependent variable ITC_{word} in the structured condition alone, but this effect did not receive substantial evidence in interaction with the random condition.

3.2. *Behavioral results*

3.2.1. *Behavioral SL outcome measures*

The rating task revealed that our participants successfully segmented the speech stream in the structured condition (see Fig. 9). Results of the CLMM analysis indicated that part-words were rated more familiar than non-words ($\beta_{part\text{-}word} = .53$, $SE = .36$, 95% CI [−.18, 1.25]), and words most familiar ($\beta_{word} = 1.68$, $SE = .43$, 95% CI [.85, 2.51]). We then evaluated with bain if: $\beta_{part\text{-}word} > 0$ & $\beta_{word} > 0$ & $\beta_{word} > \beta_{part\text{-}word}$ (section 2.5.1.). We found very strong evidence for this hypothesis ($BF_{10} = 54.63$), supported by a large Posterior Model Probability (PMP = .98). The sensitivity analysis with Bain, performed by adjusting the fraction to 2 and 3,

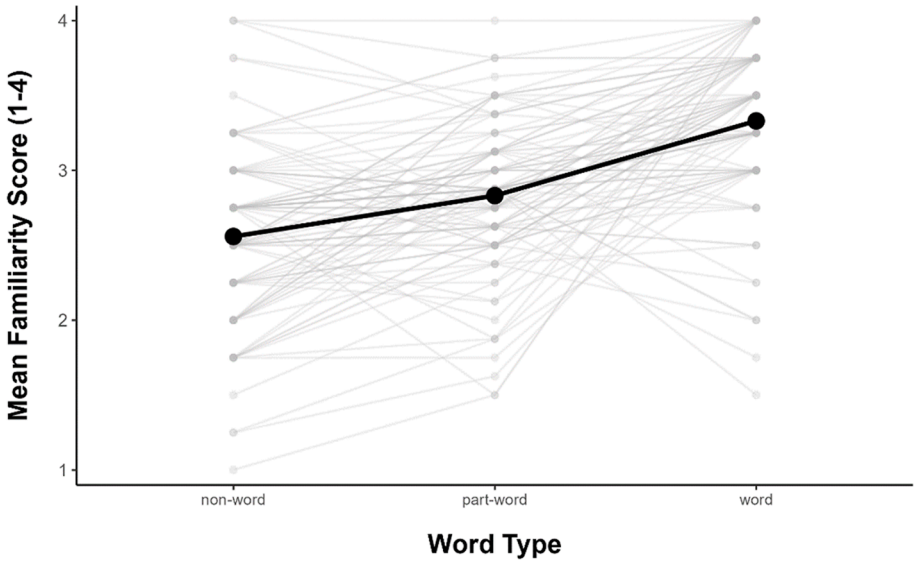

Fig. 9 – Mean familiarity ratings per word type in the rating task.
Note. **Gray dots and lines represent average rating per word type per participant. Black dots and lines represent mean ratings per word type.**

indicated a robust result ($BF_{10\ (fraction=2)} = 53.56$, PMP = .98; $BF_{10\ (fraction=3)} = 50.71$, PMP = .98). Thus, words were indeed rated most familiar compared to part-words and non-words.

With regard to the TDT, the average percentage of targets detected was 82.9%. One participant detected less than 50% (detected: 26.4%) and was therefore excluded from further analyses on this task. The effect of syllable position on decreasing RTs received extreme evidence ($\beta_{syllable\ position} = -36.57$, $SE = 1.74$, $BF_{10\ (0,\ 31.91)} > 1000$, $RR_{BF\ >\ 3}$[.09, 150]; see Fig. 10). So, participants indeed detected more predictable word-medial and -final syllables faster than initial and unpredictable syllables of the words.

3.2.2. *Brain-behavior correlations*

Contrary to our hypothesis, we found moderate evidence for a null correlation between the rating scores computed from the

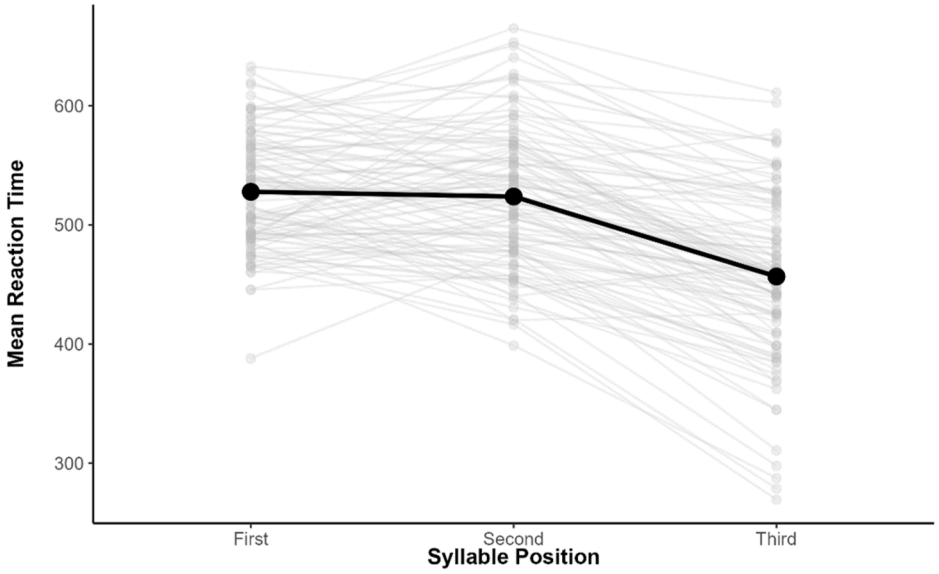

Fig. 10 – Mean reaction times per syllable position in the target detection task.
Note. **Gray dots and lines represent average RT per syllable position per participant. Black dots and lines represent mean RTs per syllable position.**

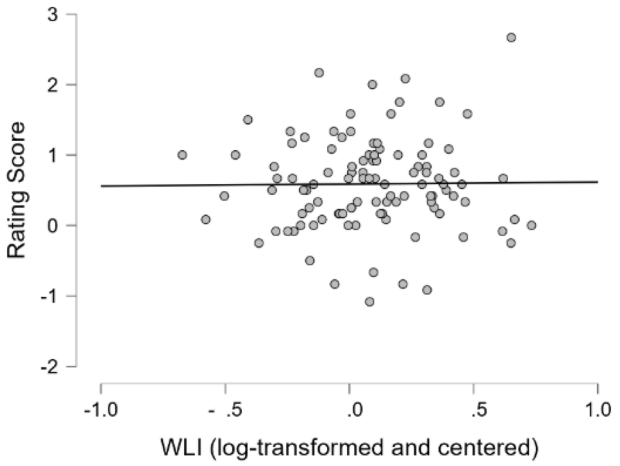

Fig. 11 – Results for the correlation between the rating score and the WLI in the structured condition.

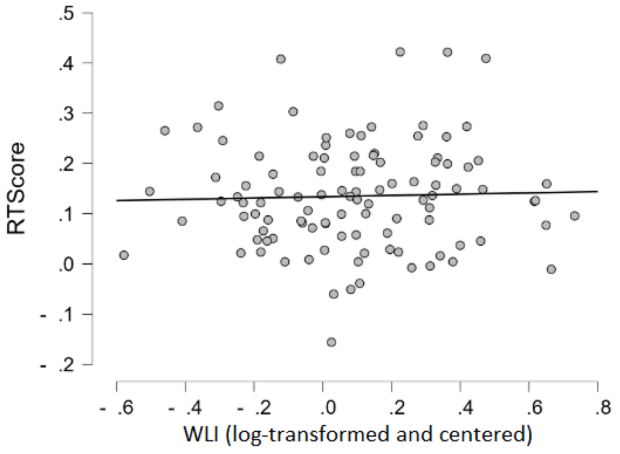

Fig. 12 – Results for the correlation between the RT facilitation scores and the WLI in the structured condition.

rating task (section 2.5.2.) and the $WLI_{structured}$ ($r = .01$, 95% CI [.00, .22], $BF_{10\ (\kappa=.5)} = .20$; Fig. 11).

With regard to the RT facilitation scores computed from the TDT, the result of the correlation with the $WLI_{structured}$ also indicated evidence for the null hypothesis ($r = .03$, 95% CI [.00, .24], $BF_{10\ (\kappa=.5)} = .24$; Fig. 12). Since the rating score and RT facilitation score both did not show evidence for a correlation with the WLI, we additionally correlated them with each other under the same prior. The rating score and RT facilitation score did correlate highly with each other (See Fig. 13; $r = .38$, 95% CI [.19, .52], $BF_{10\ (\kappa=.5)} = 790.89$).

Several studies have used the final or maximum ITC value per participant as a neural outcome measure of SL, rather than the averaged entrainment response across exposure (e.g., Choi et al., 2020; Zhang et al., 2021). In addition, our findings in section 3.1 pointed out that not the WLI, but the ITC_{word} increased over time during the structured condition. We therefore exploratively calculated the maximal ITC_{word} ($maxITC_{word}$) across bundles per participant as a new (not-preregistered) dependent variable for follow-up analyses. We correlated the $maxITC_{word}$ with the behavioral measures of SL. The $maxITC_{word}$ was not normally distributed.[17] Therefore, we decided to calculate Kendall's tau-b (τb) correlation coefficients.[18] Results revealed that the $maxITC_{word}$ correlated positively with both the rating score ($\tau b = .16$, 95% CI [.04, .28], $BF_{10\ (\kappa=.5)} = 7.18$; Fig. 14) and RT facilitation score ($\tau b = .19$, 95% CI [.06, .30], $BF_{10\ (\kappa=.5)} = 18.54$; Fig. 15), indicating that the maximal neural entrainment to the words is correlated with behavioral measures of SL, while the average WLI across exposure is not.

3.3. Correlations between tasks measuring individual differences

Given that the SSS task, CA-BAT, and PROMS were all hypothesized to measure rhythmic ability, we adhered to the preregistered prior of $\kappa = .75$ when correlating these tasks, which is suitable for larger effect sizes. For correlations between the other tasks, we adhered to the preregistered prior $\kappa = .5$ (see section 2.6). Multiple tasks measuring individual differences were not normally distributed (PPVT, CA-BAT, Digit Span, SSS task). Therefore, we calculated Kendall's τb correlation coefficients as in 3.2.2. Table 1 displays the preregistered correlations between all tasks measuring individual differences. With regard to rhythmic ability, we found very strong evidence for a positive correlation between the SSS task and the CA-BAT. We found inconclusive evidence for a correlation between the SSS task and the PROMS. Finally, we found extreme evidence for a correlation between the CA-BAT and PROMS, albeit after our last sample size update (see the sequential analysis plot in Fig. 16).

With regard to the other tasks measuring individual differences, we found evidence for positive correlations between the Gold-MSI and the PPVT, SSS task, CA-BAT, and PROMS. The PPVT showed a positive correlation with the SSS task as well. The Digit Span correlated positively with both the CA-BAT and PROMS. Sensitivity analyses indicated that these results were all robust to prior variations.[19] The PPVT and Digit Span showed moderate evidence for a correlation, but this was less robust against variations of the prior (see Fig. 17). Finally, we found evidence for no correlation ($1/6 < BF_{10} < 1/3$) between the SSS task and the Digit Span. See also Fig. 19 for a visual representation of these results.

3.4. Results of analyses investigating individual differences in statistical learning

As described in 2.6 and visualized in Fig. 5, we had planned to do a mediation analysis. We expected a direct effect of the

[17] See the JASP supplement for Shapiro–Wilk results and Q–Q plots.

[18] Kendall's τb is the non-parametric option for Bayesian correlations in JASP. A conversion table of effect sizes from Pearson's r to τb is provided by Gilpin (1993). Small effect sizes ($r = .10$–$.30$) correspond to a τb of .06–.19. Medium effect sizes ($r = .30$–$.50$) correspond to $\tau b = .20$–$.33$ and large effects ($r \geq .50$) correspond to $\tau b \geq .34$.

[19] See JASP supplement: https://osf.io/c63u8.

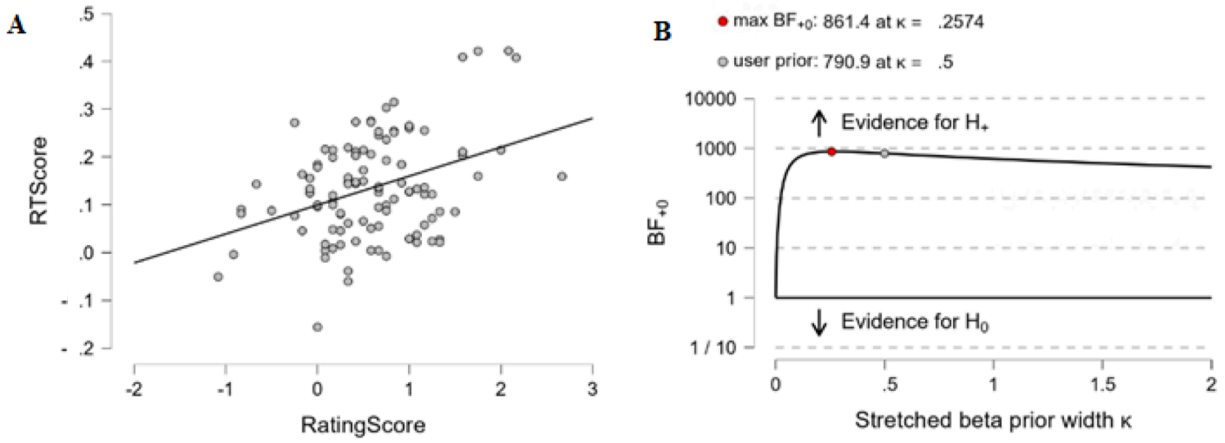

Fig. 13 – Results for the correlation between the RT facilitation score and the rating score.
***Note.* A) Scatterplot of the correlation (Pearson's r); B) Sensitivity analysis with the Bayes Factor Robustness Check, showing the BF as a function of the possible values for prior κ. Figures from JASP.**

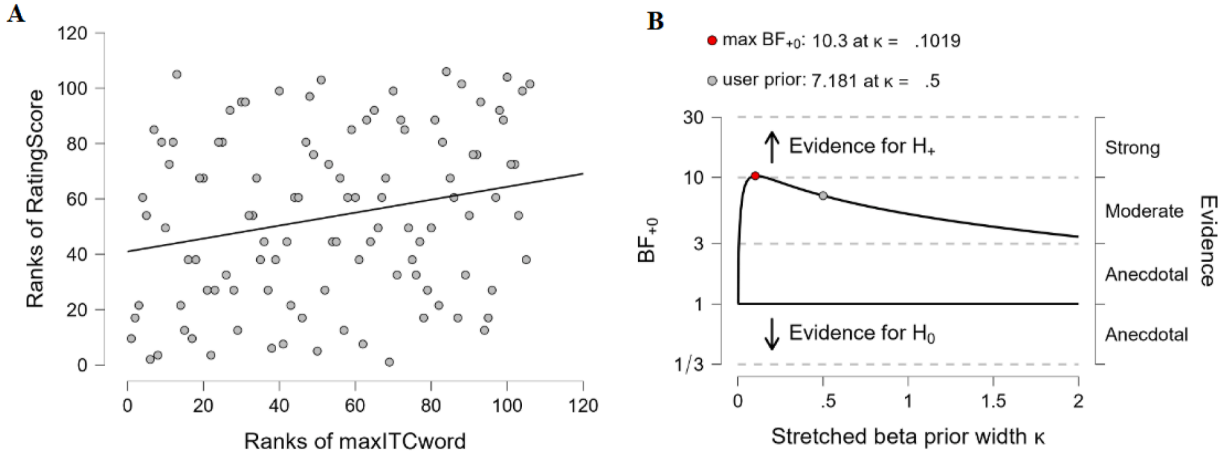

Fig. 14 – Results for the correlation between the rating score and $maxITC_{word}$.
***Note.* A) Scatterplot of the correlation (Kendall's *τb*); B) Sensitivity analysis with the Bayes Factor Robustness Check, showing the BF as a function of the possible values for prior κ. Figures from JASP.**

SSS task (PLV) on the $WLI_{structured}$, mediated by rhythmic ability. Therefore, we first tested for this preregistered direct effect, using a linear regression in Bain. Contrary to our expectation, we found evidence for the absence of a relationship between the SSS task and the $WLI_{structured}$ ($\beta = .001$, $SE = .03$, 95% CI [−.05, .06], $BF_{c\text{-}path\ =\ 0\ (fraction\ =\ 1)} = 10.14$, PMP = .91; $BF_{c\text{-}path>0\ (fraction\ =\ 1)} = 1.08$, PMP = .09). Sensitivity analyses indicated that this effect was robust ($BF_{c\text{-}path\ =\ 0\ (fraction\ =\ 2)} = 7.17$, PMP = .87; $BF_{c\text{-}path>0\ (fraction\ =\ 2)} = 1.08$, PMP = .13; $BF_{c\text{-}path\ =\ 0\ (fraction\ =\ 3)} = 5.85$, PMP = .85; $BF_{c\text{-}path>0\ (fraction\ =\ 3)} = 1.08$, PMP = .15).

Since we found evidence against a direct effect of the SSS task on the $WLI_{structured}$, we correlated all other tasks separately with the $WLI_{structured}$ in JASP under the prior $\kappa = .5$, as preregistered and described in section 2.6. The results of these correlations are displayed in Table 2. We did not find evidence for any correlations between the $WLI_{structured}$ and our tasks for individual differences. In contrast, we found moderate evidence that there was no correlation between the $WLI_{structured}$ and the CA-BAT, and robust evidence for H0 regarding the correlation between the $WLI_{structured}$ and the PROMS. Evidence for correlations between the $WLI_{structured}$ and the Gold-MSI and Digit Span was inconclusive (BF_{10} around 1). Only the correlation between the $WLI_{structured}$ and PPVT indicated moderate evidence for H1 ($3 < BF_{10} < 6$).

As we did not find evidence for correlations between any of our measures of individual differences and the $WLI_{structured}$, we explored whether these tasks would be related to the $maxITC_{word}$ (cf. section 3.2.2.), the rating score, and/or RT facilitation score. Results of these not-preregistered analyses

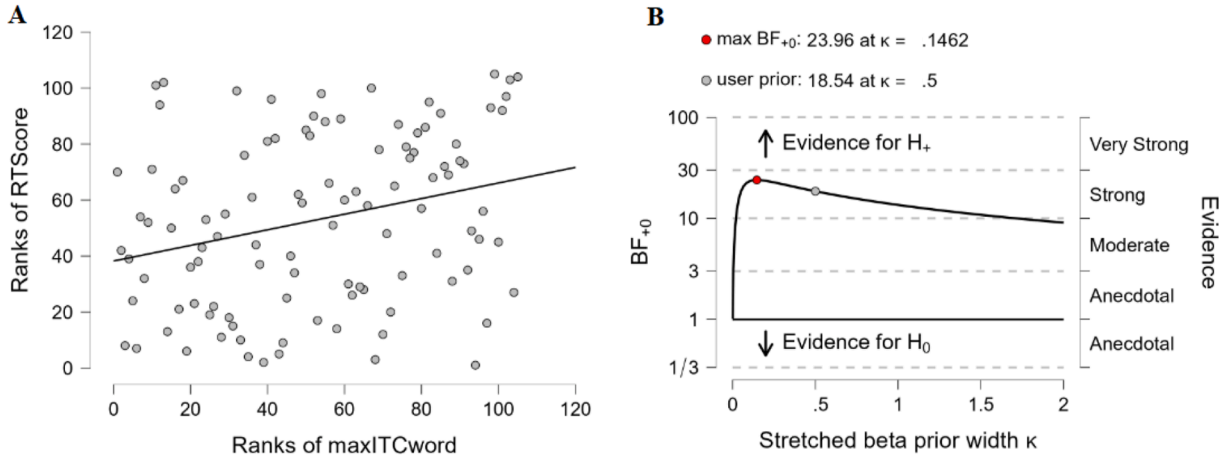

Fig. 15 – Results for the correlation between the RT facilitation score and maxITC$_{word}$.
***Note.* A) Scatterplot of the correlation (Kendall's *τb*); B) Sensitivity analysis with the Bayes Factor Robustness Check, showing the BF as a function of the possible values for prior κ. Figures from JASP.**

Table 1 – Results of the one-sided correlation analyses between all tasks measuring individual differences.

	SSS task	CA-BAT	PROMS	Gold-MSI	PPVT
CA-BAT	$N = 103$ $\tau b = .22$[b] [.09, .34] $BF_{10} = 71.26$	–			
PROMS	$N = 103$ $\tau b = .13$ [.02, .25] $BF_{10} = 1.71$	$N = 105$ $\tau b = .24$[c] [.11, .36] $BF_{10} = 204.45$	–		
Gold-MSI	$N = 103$ $\tau b = .35$[d] [.21, .46] $BF_{10} > 1000$	$N = 105$ $\tau b = .23$[c] [.09, .34] $BF_{10} = 127.37$	$N = 105$ $\tau b = .20$[b] [.07, .32] $BF_{10} = 29.99$	–	
PPVT	$N = 103$ $\tau b = .18$[b] [.05, .29] $BF_{10} = 10.79$	$N = 105$ $\tau b = .10$ [.01, .22] $BF_{10} = 1.01$	$N = 105$ $\tau b = .08$ [.01, .20] $BF_{10} = .65$	$N = 105$ $\tau b = .18$[b] [.05, .30] $BF_{10} = 16.13$	–
Digit span	$N = 103$ $\tau b = .01$[a] [.00, .15] $BF_{10} = .21$	$N = 105$ $\tau b = .16$[b] [.04, .28] $BF_{10} = 6.71$	$N = 105$ $\tau b = .20$[b] [.06, .31] $BF_{10} = 27.69$	$N = 105$ $\tau b = .09$ [.01, .21] $BF_{10} = .78$	$N = 105$ $\tau b = .14$[a] [.03, .26] $BF_{10} = 3.47$

Note. Sample sizes vary due to missing data (see 2.1.2). Correlations between the SSS task, CA-BAT, and PROMS were calculated under the prior κ = .75. All other correlations were calculated under the prior κ = .5. Values between brackets refer to the lower and upper limits of the 95% Credible Interval.
[a] $BF_{10} > 3$ or $BF_{10} < 1/3$.
[b] $BF_{10} > 6$.
[c] $BF_{10} > 100$.
[d] $BF_{10} > 1000$.

can be found in Table 3. With regard to the maxITC$_{word}$, we found moderate evidence ($1/6 < BF_{10} < 1/3$) for no correlation with the PROMS, and close-to-moderate evidence for positive correlations with the SSS task and Digit Span. With regard to the rating score, we found evidence for null correlations with the PPVT ($BF_{10} < 1/6$) and with the CA-BAT and SSS task ($BF_{10} < 1/3$). The RT facilitation score showed robust evidence for a positive correlation with the PROMS (see Fig. 18). Fig. 19 shows an overview of all correlations between the tasks used in the current study.

4. Discussion

This study aimed to uncover underpinnings of individual differences in auditory SL for word segmentation. A large sample of 106 participants performed a speech segmentation SL task while we measured their neural entrainment to the frequencies of the words and syllables with EEG. SL performance was additionally assessed through two behavioral tasks: a familiarity rating task and target detection task (TDT).

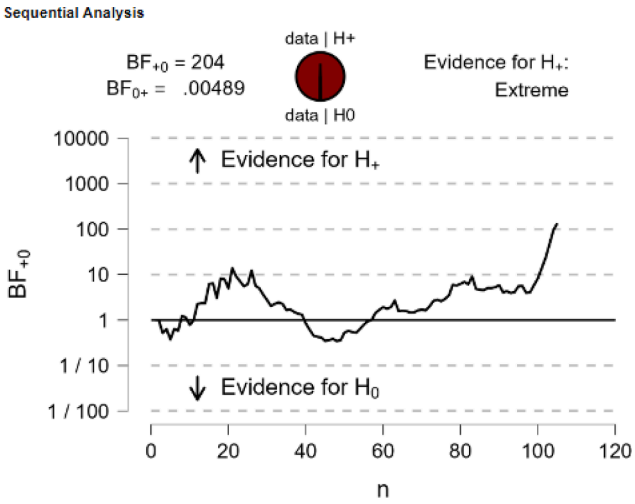

Fig. 16 – Sequential analysis plot for the correlation between the CA-BAT and PROMS under κ = .75.
***Note.* This figure shows that the BF_{10} was only larger than 6 in our last sample size update (updating with 15 participants from $N = 90$ to $N = 105$).**

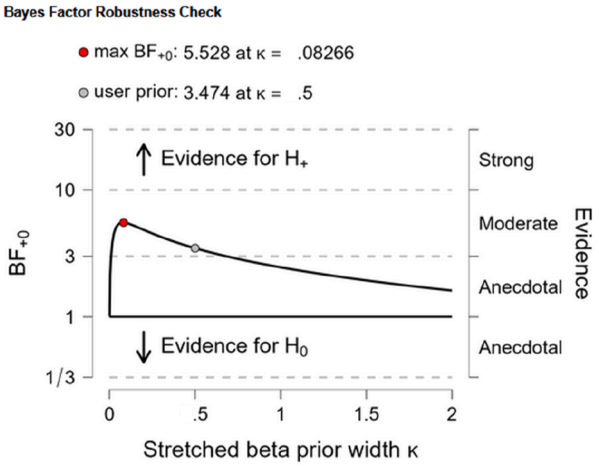

Fig. 17 – Results of the sensitivity analysis for the correlation between the Digit Span and the PPVT.
***Note.* Sensitivity analysis with the Bayes Factor Robustness Check, showing the BF as a function of the possible values for prior κ. Figure from JASP.**

Participants were further presented with a control condition consisting of randomly shuffled syllables, where word segmentation based on TPs was not possible. Finally, participants completed multiple tasks assessing musical, rhythmic, and cognitive abilities with the aim to uncover correlations between individual differences on these tasks and individual differences in SL.

4.1. *Replication of Batterink and Paller (2017)*

First, we aimed to replicate previous work on statistical learning, viz. the study by Batterink and Paller (2017), which showed neural entrainment evidence for TP-based word segmentation, as indicated by a difference in WLI between a structured and a random condition. On the one hand, we have replicated this effect in our study, indicating that neural entrainment is a reliable measure of SL. On the other hand, we found different results regarding the time-course of learning. In contrast to the earlier finding by Batterink and Paller (2017), the WLI in the current study did not show statistical evidence for different trajectories over time between conditions, as indicated by evidence for the null hypothesis. However, this finding may be attributed to the absence of a decreasing $ITC_{syllable}$ in our data, as our participants showed a relatively stable $ITC_{syllable}$ across time. Batterink and Paller (2017) found both a decrease of $ITC_{syllable}$ and an increase of ITC_{word} in the structured condition. The WLI is a composite of these two ITC measures (see 1.2), which increases both when ITC_{word}

Table 2 – Correlation analyses between the $WLI_{structured}$ and tasks for individual differences.

	CA-BAT	PROMS	Gold-MSI	Digit Span	PPVT
$WLI_{structured}$	$N = 105$ $rb = -.01$[a] [.00, .14] $BF_{10} = .18$	$N = 105$ $rb = -.01$[b] [.00, .14] $BF_{10} = .165$	$N = 106$ $rb = .09$ [.01, .21] $BF_{10} = .84$	$N = 105$ $rb = .10$ [.01, .22] $BF_{10} = 1.03$	$N = 105$ $rb = .15$[a] [.03, .27] $BF_{10} = 4.44$

Note. Prior was $\kappa = .5$.
Values between brackets refer to the lower and upper limits of the 95% Credible Interval.
[a] $BF_{10} < 1/3$ or $BF_{10} > 3$.
[b] $BF_{10} < 1/6$ or $BF_{10} > 6$.

Table 3 – Correlation analyses between the $maxITC_{word}$, rating score and RT facilitation score, with tasks for individual differences.

	CA-BAT	PROMS	Gold-MSI	SSS task	Digit Span	PPVT
$maxITC_{word}$	$N = 105$ $rb = .07$ [.01, .20] $BF_{10} = .52$	$N = 105$ $rb = .03$[a] [.00, .17] $BF_{10} = .29$	$N = 106$ $rb = .08$ [.01, .20] $BF_{10} = .65$	$N = 103$ $rb = .14$ [.02, .26] $BF_{10} = 2.78$	$N = 105$ $rb = .13$ [.02, .25] $BF_{10} = 2.29$	$N = 105$ $rb = .10$ [.01, .23] $BF_{10} = 1.16$
Rating score	$N = 105$ $rb = .03$[a] [.00, .17] $BF_{10} = .27$	$N = 105$ $rb = .09$ [.01, .21] $BF_{10} = .84$	$N = 106$ $rb = .10$ [.01, .23] $BF_{10} = 1.20$	$N = 103$ $rb = .01$[a] [.00, .15] $BF_{10} = .20$	$N = 105$ $rb = .05$ [.00, .18] $BF_{10} = .38$	$N = 105$ $rb = -.02$[b] [.00, .14] $BF_{10} = .16$
RT facilitation score	$N = 104$ $rb = .04$ [.00, .18] $BF_{10} = .34$	$N = 104$ $rb = .16$[b] [.04, .28] $BF_{10} = 7.11$	$N = 105$ $rb = .08$ [.01, .21] $BF_{10} = .75$	$N = 102$ $rb = .06$ [.01, .19] $BF_{10} = .43$	$N = 104$ $rb = .06$ [.01, .19] $BF_{10} = .44$	$N = 104$ $rb = .06$ [.01, .19] $BF_{10} = .48$

Note. Prior was $\kappa = .5$.
Values between brackets refer to the lower and upper limits of the 95% Credible Interval.
[a] $BF_{10} < 1/3$ or $BF_{10} > 3$.
[b] $BF_{10} < 1/6$ or $BF_{10} > 6$.

increases, but also when $ITC_{syllable}$ decreases. When we performed a not-preregistered follow-up analysis in which we examined the time-course of learning by focusing on the ITC_{word} alone as the dependent variable in the structured condition, we did find evidence for the alternative hypothesis that entrainment at the word level increased over time. This result could inspire future studies to independently look at ITC_{word} and $ITC_{syllable}$, where ITC_{word} is taken to be the measure of SL instead of a composite variable such as the WLI. Separate consideration of ITC_{word} and $ITC_{syllable}$ has already been adopted in several studies following Batterink and Paller (2017)'s initial study (e.g., Batterink & Paller, 2019; Moreau et al., 2022; Pinto et al., 2022; Zhang et al., 2021).

We furthermore replicated the behavioral results of Batterink and Paller (2017) regarding the rating task and TDT, with results of preregistered analyses regarding both tasks providing evidence that our participants became sensitive to the statistical regularities in the structured stream. Performance on these tasks was also positively correlated. Furthermore, we tested as preregistered if the WLI in the structured condition correlated with performance on these behavioral tasks. To our surprise, this analysis yielded evidence for the null hypothesis. However, as discussed above, the WLI may not be the most sensitive measure for SL in our data. We therefore considered an alternative not-preregistered dependent variable as an individual neural index of SL: the $maxITC_{word}$ from the time-course bundle-based analysis in the structured condition. The $maxITC_{word}$ represents the highest ITC_{word} for each individual participant across exposure and may reflect each participant's peak sensitivity to the statistical structure. Interestingly, the $maxITC_{word}$ did correlate positively with both behavioral measures of SL.

Participants' sensitivity to the structure likely waxes and wanes over time, due to the length of the exposure period (Henry & Herrmann, 2014). An individual's peak sensitivity to the statistical properties of the speech stream (in other words, the moment the participant has fully recognized the TP-structure in the input) may therefore be a *more* relevant indication of learning outcomes, rather than their average sensitivity over time. It is possible that after this peak, participants start focusing their attention to other properties of the input stream, and that neural entrainment to the ITC_{word} diminishes as a result of this diverted attention. Batterink and Paller (2017) did not directly compare $maxITC_{word}$ with the WLI, and it is possible that the $maxITC_{word}$ is a more sensitive individual neural marker of learning than measures that are aggregated across exposure. Future studies may wish to incorporate neural indices that capture peak entrainment to words over the period of learning, such as the $maxITC_{word}$.

4.1.1. *Stimulus properties driving time-course of learning*

In the current study, ITC_{word} did not show an increase until relatively late during exposure to the structured stream (Fig. 8A), while $ITC_{syllable}$ remained relatively stable (Fig. 8B) and did not decrease as it did in Batterink and Paller (2017). A late increase in ITC_{word} was also previously found in a group of adults with dyslexia in a study by Zhang et al. (2021),

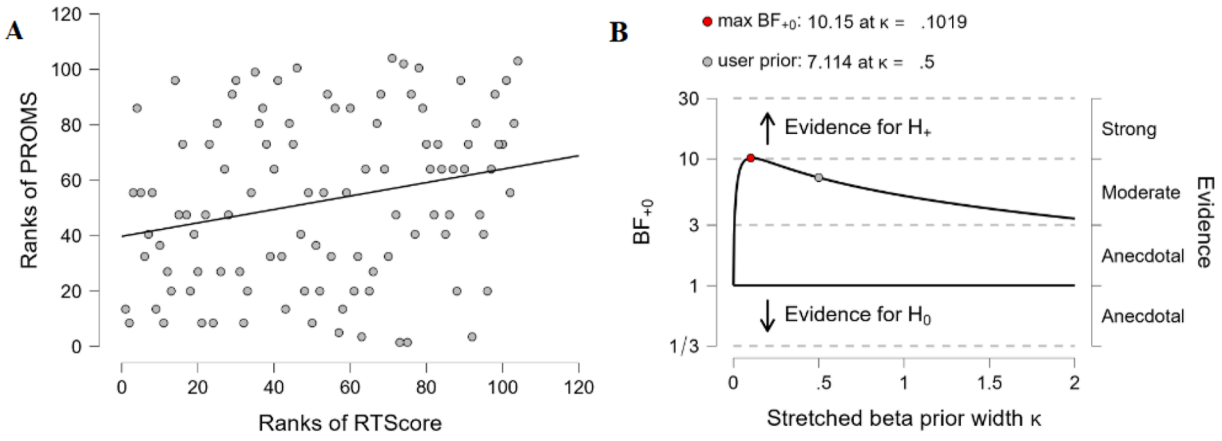

Fig. 18 – Results for the correlation between the RT facilitation score and the PROMS.
***Note.* A) Scatterplot of the correlation (Kendall's *τb*); B) Sensitivity analysis with the Bayes Factor Robustness Check, showing the BF as a function of the possible values for prior κ. Figures from JASP.**

compared to typical readers showing an earlier increase in $ITC_{word,}$ followed by a decrease. This result suggests that the overall difficulty level or learning challenge faced by an individual learner may influence their temporal trajectory of learning. While the current study did not include any adults diagnosed with dyslexia (see section 2.1.2.), it differed from previous work in terms of the stimuli. Not only did we create entirely new stimuli, but we also made sure that these speech streams were coarticulated to resemble natural continuous speech more closely. Perhaps this coarticulation made it more difficult to parse the speech stream into words than when individually recorded syllables are concatenated, resulting in delayed learning.

In addition, the overall typicality or familiarity of the syllables themselves may also have influenced the time course of neural entrainment at both word and syllable frequencies. In the current study, we avoided using syllables that were existing single-syllabic words or frequent forms in Dutch, and therefore selected relatively infrequent syllables in the Dutch language (see 2.2). This factor was not controlled for in the stimuli previously employed by Batterink and Paller (2017) and related work (e.g., Saffran, Aslin et al., 1996), and in fact, many of the syllables used in these studies were identical to existing English words (e.g., "go", "to", "row", etc.). Previous work has shown that statistical learning operates more efficiently across syllables that are commonly found in a participant's native language, compared to syllables that are rarely found (Ordin et al., 2021). Perhaps statistical learning occurs more slowly over less familiar syllables as first a representation for each individual syllable must be created, followed by concatenation into trisyllabic items. This may have slowed down the time course of learning, at least as measured by ITC_{word}, though our participants did still learn well eventually as indicated by statistical evidence for the overall entrainment difference between conditions and robust performance on the rating task and TDT. It may also provide an explanation for the relatively stable $ITC_{syllable}$, indicating that our participants were paying relatively constant attention to the less familiar syllables instead of showing a decrease in $ITC_{syllable}$ as a result of habituation.

4.2. *Investigating individual differences in statistical learning*

4.2.1. *Relations between measures of individual differences*

As illustrated in Fig. 19, we found statistical evidence for preregistered analyses indicating multiple positive correlations between performance on our measures of individual differences (see 2.3.3. for detailed descriptions of the tasks). Specifically, we were interested in whether the tasks aiming to assess rhythmic ability were correlated. A preregistered analysis indeed supplied evidence for a positive correlation between the CA-BAT and the PROMS, but only after our final sample size update (section 3.3; Fig. 16). This is probably due to the different approaches these tasks take to measuring rhythmic ability: the CA-BAT focuses on judgements of beat alignment in which two tracks with metronome beeps over naturalistic music are compared, whereas the PROMS relies more on memory as it requires participants to remember a rhythm sequence and compare it to another one. This may also explain why the SSS task correlated positively with the CA-BAT, but the correlation analysis regarding the SSS task and PROMS yielded inconclusive evidence. The SSS task measures whether participants synchronize their whispering to the auditory input: both the SSS and CA-BAT tasks share (perceived or produced) synchronization ability as a common factor.

Furthermore, the Gold-MSI questionnaire gauging general musicality including musical training experience correlated with all three rhythm-related tasks, indicating that self-reported musicality relates to actual performance on these laboratory tasks. In particular, we found extreme evidence

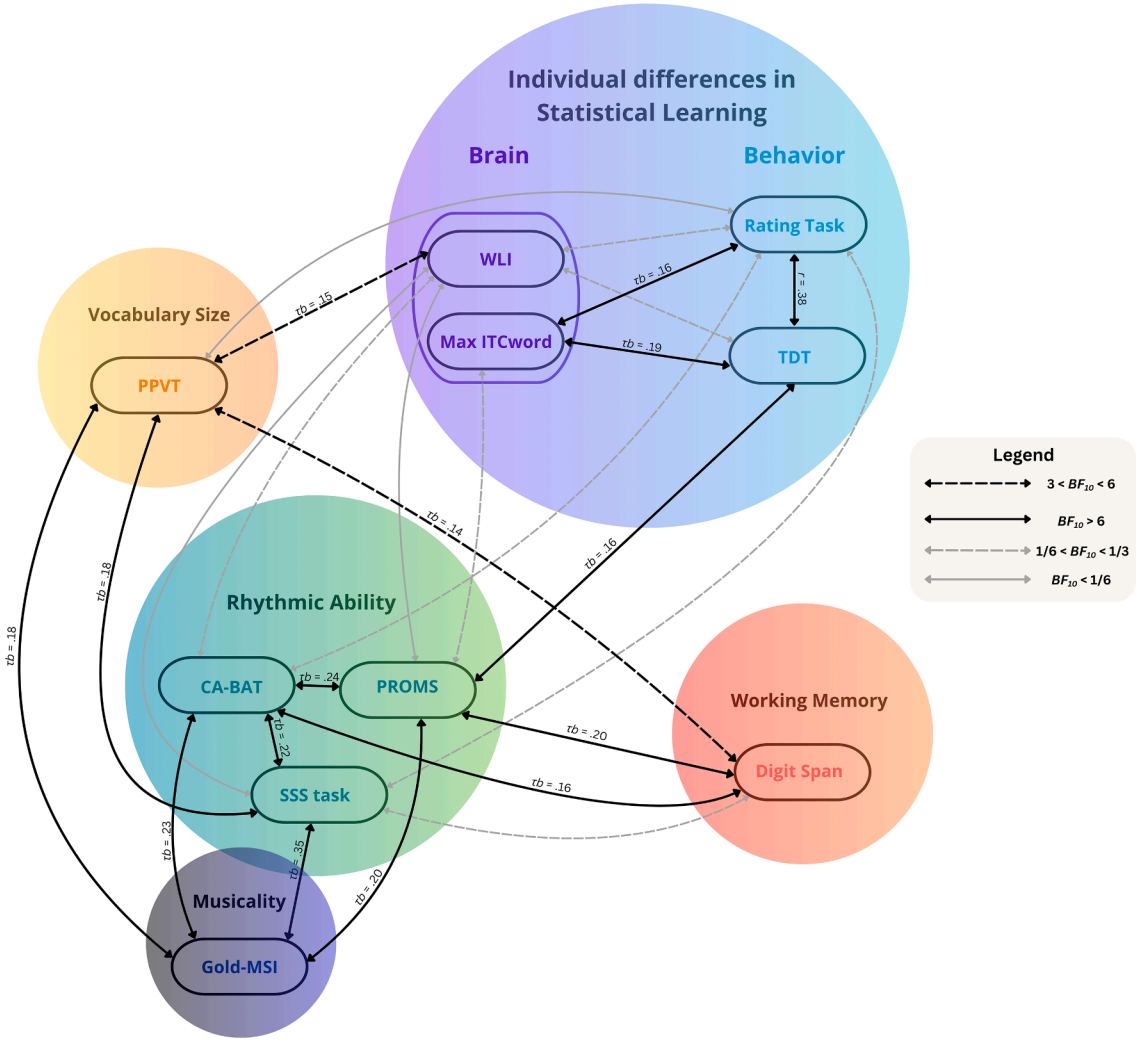

Fig. 19 – Figure indicating positive and null correlations found between the tasks in the current study. *Note*. Effect sizes are provided for all correlations that are not null. Black lines indicate evidence for correlations, grey lines indicate evidence for null effects. Small effect sizes ($r = .10 - .30$) correspond to a τb of .06 – .19. Medium effect sizes ($r = .30 - .50$) correspond to $\tau b = .20 - .33$ and large effects ($r \geq .50$) correspond to $\tau b \geq .34$.

($BF_{10} > 1000$) for a positive correlation between the SSS task and the Gold-MSI. This result, together with the SSS task and the CA-BAT correlation, indicates that the SSS task is highly related to musicality including musical training experience, and musical rhythmic ability.

We additionally found evidence for positive relationships between both the CA-BAT and PROMS with the Digit Span, assessing working memory. Both these tasks involve listening to multiple sound excerpts and answering a question about these afterwards. However, the PROMS likely relies more on working memory than the CA-BAT (also indicated by the larger effect size; see Table 1 and Fig. 19), since in principle a correct response to an item in the CA-BAT could be based on just one of the two musical excerpts. For example, if the participant perceives the second music excerpts to have beats unaligned with the rhythm of the music, the participant can confidently indicate that the first excerpt was correct – even when the participant had forgotten the first excerpt. The PROMS, on the other hand, requires the participant to compare two rhythmic sequences (i.e., state whether they were identical or not). If the participant forgot either of the sequences, they would not be confident about whether the rhythms were the same. The task included graded answers ("definitely different/the same" *vs* "probably different/the same," as well as an "I don't know" option), but maximal points on the task are only obtained when the most extreme answers are chosen. Finally, auditory-motor synchronization did not seem to rely on working memory, provided by the moderate evidence for no relation between the SSS task and the Digit Span. This task relied on sub-conscious rhythmic production, being the most 'online' task of rhythmic ability out of these three tasks.

Finally, vocabulary size, as measured with the PPVT, showed evidence for a correlation with both the SSS task and the Gold-MSI (Table 1; Fig. 19). These results are in line with findings that musicality positively influences linguistic ability

including vocabulary size (e.g., Ladányi et al., 2020 (review); Zuk et al., 2022). However, the PPVT included items regarding musical terminology, and as the SSS task and Gold-MSI were highly correlated, musical training may have influenced these results.

4.2.2. *No effect of SSS task performance on statistical learning*

The main goal of the present study was to investigate individual differences in SL. We hypothesized that individuals with better musical – specifically rhythmic – abilities would show better SL in the context of speech segmentation. We had operationalized this relationship in our preregistration as a direct effect of the SSS task on the $WLI_{structured}$, mediated by musical rhythmic ability (Fig. 5). This hypothesis was based on previous findings by Assaneo et al. (2019) indicating that 'high synchronizers' (i.e., participants with a higher *phase-locking value* (PLV) on the SSS task) performed better on a recognition-based SL task than 'low synchronizers' (section 1.4.). Furthermore, these high synchronizers showed more white matter integrity in the dorsal language stream (Hickok & Poeppel, 2007). We connected these findings to a hypothesis by François et al., 2012, stating that the dorsal pathway could be improved in musically trained individuals and that this in turn would benefit SL (see section 1.5). As Assaneo et al. (2019) argued that musical training did not explain their data more than high/low synchronizer status, we hypothesized that in some individuals the dorsal stream would be organized more efficiently as part of the neurological substrate of innate musical ability.

However, the results from the current study supported the null hypothesis that there was no effect of SSS performance on the WLI, contrary to our hypothesis and the results of Assaneo et al. (2019). When we performed a not-preregistered follow-up analysis testing for this relation with the $maxITC_{word}$ (Table 3), the effect size was small and yielded inconclusive evidence. In both cases, we directly regressed the SSS PLV on a neural outcome measure of SL, hypothesizing this to be more sensitive to finding a relationship between these continuous variables than dividing the participants into groups. We followed up (not-preregistered) by also correlating the SSS PLV with our behavioral measures of SL (Table 3) and found evidence that there was no correlation with performance on the rating task, and inconclusive evidence in the direction of evidence for the null regarding the TDT ($BF_{10} = .43$; Table 3). Specifically, the rating task conceptually fails to replicate the findings by Assaneo et al. (2019), since it is an explicit measure of SL similar to the two-alternative forced choice task they employed. Finally, we performed a not-preregistered exploratory analysis in which we followed the protocol in Lizcano-Cortés et al. (2022) to divide our participants into high and low synchronizer groups as well, which did not alter any of these results (see Supplementary Materials S.2.)[20]. Thus, the current study provides substantial evidence that the SSS task is related to musical and rhythmic ability (see also 4.2.1.), but does not relate to individual differences in linguistic SL.

4.2.3. *No relation between measures of rhythmic ability and online statistical learning*

We hypothesized that rhythmic ability specifically would predict individual differences in SL. This hypothesis was based on literature showing that precise phase-locking of neural oscillations to auditory stimuli reflects optimal processing (e.g., Assaneo et al., 2019; Peelle & Davis, 2012; Poeppel & Assaneo, 2020). Furthermore, several previous studies reported correlations between musicality and SL (François et al., 2012; Mandikal Vasuki et al., 2017; François & Schön, 2011; Shook et al., 2013). We predicted that efficient brain-stimulus phase-locking would be supported by rhythmic abilities relevant for both music and language processing and would be reflected by stronger neural entrainment during SL (section 1.5). Therefore, we hypothesized that rhythmic ability could be a mechanism supporting SL for speech segmentation.

In contrast to our expectation, we did not obtain evidence for the preregistered analyses correlating performance on the CA-BAT, PROMS, or Gold-MSI with the $WLI_{structured}$. Our Bayesian analyses indicated evidence for the absence of such effects for the CA-BAT and PROMS, and inconclusive evidence for the Gold-MSI (see Table 2 and Fig. 19). We also found null results in our not-preregistered follow-up analyses taking the $maxITC_{word}$ as the dependent variable, yielding evidence for the null hypothesis regarding the PROMS, and inconclusive evidence regarding the CA-BAT and Gold-MSI (Table 3). The sole positive correlation between any of our rhythmic ability measures and any SL measures that received evidence for the alternative hypothesis was between the PROMS and the RT facilitation score computed from the TDT. This is surprising, as the PROMS showed inconclusive evidence for a correlation with the SSS task – which was initially hypothesized to relate to SL – and the PROMS showed evidence that it was positively correlated with the CA-BAT only after the final sample size update (Fig. 16). Furthermore, performance on the PROMS was strongly associated with working memory capacity as measured by the Digit Span (section 4.2.1.). Perhaps this relation between the RT facilitation score and the PROMS reflects a commonality in accurate memory for auditory sequences more so than rhythmic abilities per se.

Taken together, the current results suggest that musical and rhythmic abilities do not relate to individual differences in SL, which contrasts with previous literature reporting correlations between musicality and SL However, these previous studies did not use speech stimuli, but instead assessed learning of regularities from sung languages (François et al., 2012; François & Schön, 2011), pure tones (Mandikal Vasuki et al., 2017), or Morse codes (Shook et al., 2013). To our knowledge, no study has explicitly made a connection between musical ability and SL of speech. Thus, a possible explanation for our results contrasting with these findings could be that musicality influences SL in a highly domain-specific way, rather than universally influencing all types of SL. Previous work shows that SL of speech is highly domain-specific (Siegelman & Frost, 2015), and particularly SL of speech is entrenched by prior linguistic knowledge (Siegelman et al., 2018).

Furthermore, while we assessed rhythmic ability through multiple tasks, they were all behavioral in nature and relied

[20] https://osf.io/vf4qj.

either on explicit judgements or sensorimotor synchronization. We had hypothesized that rhythmic ability may be related to SL by sharing a common neural substrate and leading to more efficient neural processing of auditory stimuli. Whereas we measured SL as 'directly' as possible with neural entrainment, this was not the case for rhythmic ability. Conceivably, using neural measures for both rhythmic ability and SL could reveal positive correlations, in line with our initial hypothesis. In fact, this possibility is supported by our recent study in infants (van der Wulp et al., 2025). In this study, we found that infants showing stronger neural entrainment to the frequency of the meter in an auditorily presented rhythm also showed stronger entrainment to the frequency of words in a structured stream identical to the one used in the current study. Conducting a similar study in adults would be a promising avenue for further research.

Another explanation for our results may be that our sample of typically developed adults is sufficiently equipped to perform SL, and thus there is not enough variation in SL performance to correlate with other cognitive differences, such as differences in rhythmic ability. As the literature points out, individuals with language impairments in particular appear to show impaired musical and specifically rhythmic abilities (e.g., Boll-Avetisyan et al., 2020; Caccia & Lorusso, 2020; Fiveash et al., 2021; Flaugnacco et al., 2014; Huss et al., 2011; Kraus et al., 2014; Ladányi et al., 2020; Sallat & Jentschke, 2015). Further investigations using an individual differences approach on more diverse samples, including participants with language impairments, could shed more light on this possibility.

4.2.4. *Working memory*

We broadened our search for individual differences in SL to working memory capacity by including the forward Digit Span. As earlier studies discussed in section 1.4 did not find conclusive evidence regarding a relation between working memory and SL, the question of whether working memory related to SL in our sample was preregistered as exploratory. In line with the literature, we also did not find any conclusive evidence for or against a relationship between the Digit Span and our measures of SL. Only the correlation between the not-preregistered $maxITC_{word}$ and the Digit Span received anecdotal evidence. Moreover, the effect size was quite small (Table 3). In sum, it seems that working memory has, at best, a small effect on SL. It is therefore unlikely that working memory is a driving factor of linguistic SL in the typical population.

4.2.5. *Individual differences in SL and adult vocabulary*

We administered a vocabulary test (PPVT), in order to add to the body of research in children indicating a relationship between individual differences in SL and vocabulary size. Here, vocabulary size is interpreted as an outcome measure of SL, rather than a source of individual variability in SL (see Fig. 1). Our aim was to investigate whether vocabulary and SL are also related in adulthood, or whether this is specific to children. This analysis was preregistered as exploratory. Our results suggest that the role of SL in adult vocabulary appears to be modest, as we found moderate evidence for a small correlation between the PPVT and the preregistered WLI. However, this was not the case for the (not-preregistered) $maxITC_{word}$, where the evidence was inconclusive (Table 3). We also found no evidence of a relationship with vocabulary size for our behavioral measures of SL, as indicated by evidence for the null on the rating task and inconclusive evidence on the TDT.

One possible explanation for the limited role of SL in predicting adult vocabulary attainment may be developmental changes in the importance of SL as a mechanism for vocabulary learning. It is possible that SL plays a central role in vocabulary learning early in development, but that explicit learning mechanisms contribute more to new vocabulary growth by adulthood (e.g., Batterink & Neville, 2011). In addition, environmental and socio-cultural factors are associated with adults' differential exposure to new words (e.g., educational attainment, occupation, reading preferences). This explanation is consistent with Misyak and Christiansen (2012), who found that vocabulary size in adulthood was more related to print exposure than SL. Overall, our findings suggest that adult participants' vocabulary acquisition was multifaceted, and not only predicted by SL ability.

5. Conclusions and theoretical implications

The current Registered Report aimed to investigate individual differences in SL, by replicating previous work (Assaneo et al., 2019; Batterink & Paller, 2017) and by extending it through investigating relations between rhythmic and musical ability, as well as working memory and vocabulary size to neural and behavioral measures of SL. We have indeed replicated the main effects of Batterink and Paller (2017), showing a difference between the structured and random condition in the neural measures of SL. Neural entrainment to the regularities in the structured condition increased over time, but only as measured through ITC_{word}, rather than our preregistered metric, the WLI. Furthermore, successful learning in our sample was attested through our two behavioral tasks. Interestingly, in not-preregistered follow-up analyses we found that each individual's maximal ITC_{word} robustly predicted their SL performance on both behavioral tasks. In contrast to our preregistered expectations, we found evidence for the absence of an effect of the SSS task (Assaneo et al., 2019) on SL, as well as no effects of rhythmic ability on the measurements of SL. This was indicated by either evidence for the null hypothesis or inconclusive evidence, depending on the task (see Tables 2 and 3). Evidence regarding working memory remained inconclusive. Finally, we found moderate evidence for a small correlation between vocabulary size and the WLI. However, we found evidence for the null hypothesis between the PPVT and one behavioral measure of SL (the rating task), as well as inconclusive evidence on the other (the TDT).

Overall, our results suggest that linguistic SL stands largely independently from other individual skills and aptitudes. This

is in line with the view put forth by Siegelman and Frost (2015), who argued that SL is independent of general cognitive abilities. Here, we extend this to other (musical, rhythmic) abilities beyond general intelligence and working memory. Furthermore, we investigated these possible relationships using (behavioral and neural) measures of SL beyond those that (exclusively) rely on explicit recognition abilities. In contrast, we assessed rhythmic ability through multiple behavioral tasks that all relied on either explicit judgements or sensorimotor synchronization. It may be that these behavioral measurements are not 'direct' enough, as we in the same vein have indications that the neural entrainment measure for SL is more directly indexing the identification component of SL than behavioral measures (cf. section 1.2.).

Our data indicates that there are individual differences in SL, and that performance on different measures of SL is correlated, but these individual differences do not appear to strongly relate to other individual abilities in our sample. It is possible that this general lack of correlation could be due to our sample consisting of healthy, typically developed adults. Populations outside of typical development have been found to show weaker SL (e.g., Evans et al., 2009; Gabay et al., 2015; Lammertink et al., 2017; Newman et al., 2016; Singh et al., 2012; Vandermosten et al., 2019; Zhang et al., 2021), which may lead to stronger relationships between SL and other individual abilities. We speculate that in individuals with sufficient or typical SL abilities, their SL abilities may not relate to other aspects of cognition. However, if the functioning of the normal SL capacity breaks down or is atypical, as in populations with diverse types of language disorders, a relationship may emerge. For instance, rhythmic ability is found to be impaired in populations with language impairments (e.g., Ladányi et al., 2020), so the hypothesized relation between rhythmic abilities and SL may be present in populations outside of typical development. Inclusion of broader, more diverse samples in the study of individual differences in SL represents an important direction for future work in this field.

CRediT authorship contribution statement

I.M. van der Wulp: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **M.E. Struiksma:** Writing – review & editing, Methodology, Conceptualization. **L.J. Batterink:** Writing – review & editing, Methodology, Conceptualization. **F.N.K. Wijnen:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization.

Stage 2 registered report

This is the Stage 2 Registered Report including the results and discussion of our study.

The Stage 1 report can be found here: https://osf.io/2y6sx.

Funding

This work is funded by the Netherlands Organization for Scientific Research (NWO), project number PGW.21.007.

Conflict of interest disclosure

The authors of this article declare that they have no financial conflict of interest with the content of this article.

Acknowledgements

We are very thankful to our MA students Mila Brandsen, Carmen Olsthoorn, Julia Koekkoek, and Sangyu Chen for their assistance with parts of the data collection for this project. We want to express our appreciation to the Utrecht University Institute for Language Science labs and the lab support staff for their technical and practical assistance. We would further like to thank Karin Wanrooij for her help with the creation of the stimuli, Betül Boz for her help with the practical preparations for the experiment and providing us with an additional pilot sample, as well as Kirsten Schutter and Herbert Hoijtink for their input on the statistical analyses. We would also like to thank Henkjan Honing for suggesting the CA-BAT and Gold-MSI. Finally, we would like to thank Elizabeth Wonnacott and two anonymous reviewers for their insights on earlier versions of this manuscript at Stages 1 and 2.

Scientific transparency statement

DATA: All raw and processed data supporting this research are publicly available: https://doi.org/10.24416/UU01-2GP6BV, https://doi.org/10.17605/OSF.IO/JHBE8.
CODE: All analysis code supporting this research is publicly available: https://doi.org/10.17605/OSF.IO/JHBE8, https://doi.org/10.24416/UU01-2GP6BV.
MATERIALS: All study materials supporting this research are publicly available: https://doi.org/10.17605/OSF.IO/JHBE8, https://doi.org/10.24416/UU01-2GP6BV.
DESIGN: This article reports, for all studies, how the author(s) determined all sample sizes, all data exclusions, all data inclusion and exclusion criteria, and whether inclusion and exclusion criteria were established prior to data analysis.
PRE-REGISTRATION: At least part of the study procedures was pre-registered in a time-stamped, institutional registry prior to the research being conducted: https://osf.io/2y6sx. The analyses that were undertaken deviated from the preregistered analysis plans. All such deviations are fully disclosed in the manuscript.

For full details, see the *Scientific Transparency Report* in the supplementary data to the online version of this article.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2025.09.012.

REFERENCES

Arciuli, J. (2017). The multi-component nature of statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1711), Article 20160058. https://doi.org/10.1098/rstb.2016.0058

Asano, R. (2022). The evolution of hierarchical structure building capacity for language and music: A bottom-up perspective. *Primates, 63*(5), 417–428. https://doi.org/10.1007/s10329-021-00905-x

Assaneo, M. F., Ripollés, P., Orpella, J., Lin, W. M., de Diego-Balaguer, R., & Poeppel, D. (2019). Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nature Neuroscience, 22*(4), 627–632. https://doi.org/10.1038/s41593-019-0353-z

Baguley, T., & Kaye, W. S. (2010). Review of: Understanding psychology as a science: An introduction to scientific and statistical inference, by Z. Dienes. *British Journal of Mathematical and Statistical Psychology, 63*(3), 695–698.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Batterink, L. J. (2017). Rapid statistical learning supporting word extraction from continuous speech. *Psychological Science, 28*(7), 921–928. https://doi.org/10.1177/0956797617698226

Batterink, L. J., & Choi, D. (2021). Optimizing steady-state responses to index statistical learning: Response to Benjamin and colleagues. *Cortex, 142*, 379–388. https://doi.org/10.1016/j.cortex.2021.06.008

Batterink, L., & Neville, H. (2011). Implicit and explicit mechanisms of word learning in a narrative context: An event-related potential study. *Journal of Cognitive Neuroscience, 23*(11), 3181–3196.

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

Bekius, A., Cope, T. E., & Grube, M. (2016). The beat to read: A cross-lingual link between rhythmic regularity perception and reading skill. *Frontiers in Human Neuroscience, 10*. https://www.frontiersin.org/article/10.3389/fnhum.2016.00425.

Benjamin, L., Dehaene-Lambertz, G., & Fló, A. (2021). Remarks on the analysis of steady-state responses: Spurious artifacts introduced by overlapping epochs. *Cortex, 142*, 370–378. https://doi.org/10.1016/j.cortex.2021.05.023

Bogaerts, L., Siegelman, N., Christiansen, M. H., & Frost, R. (2022). Is there such a thing as a 'good statistical learner'? *Trends in Cognitive Sciences, 26*(1), 25–37. https://doi.org/10.1016/j.tics.2021.10.012

Boll-Avetisyan, N., Bhatara, A., & Höhle, B. (2020). Processing of rhythm in speech and music in adult dyslexia. *Brain Sciences, 10*(5), 261. https://doi.org/10.3390/brainsci10050261

Bouwer, F. L., Werner, C. M., Knetemann, M., & Honing, H. (2016). Disentangling beat perception from sequential learning and examining the influence of attention and musical abilities on ERP responses to rhythm. *Neuropsychologia, 85*, 80–90. https://doi.org/10.1016/j.neuropsychologia.2016.02.018

Caccia, M., & Lorusso, M. L. (2020). The processing of rhythmic structures in music and prosody by children with developmental dyslexia and developmental language disorder. *Developmental Science.* , Article e12981. https://doi.org/10.1111/desc.12981

Choi, D., Batterink, L. J., Black, A. K., Paller, K. A., & Werker, J. F. (2020). Preverbal infants discover statistical word patterns at similar rates as adults: Evidence from neural entrainment. *Psychological Science, 31*(9), 1161–1173. https://doi.org/10.1177/0956797620933237

Christensen, R. H. B. (2022). *Ordinal—Regression models for ordinal data." R package version 2022* (pp. 11–16). https://CRAN.R-project.org/package=ordinal.

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition, 114*(3), 356–371. https://doi.org/10.1016/j.cognition.2009.10.009

Daikoku, T., & Goswami, U. (2022). Hierarchical amplitude modulation structures and rhythm patterns: Comparing Western musical genres, song, and nature sounds to Babytalk. *Plos One, 17*(10), Article e0275631. https://doi.org/10.1371/journal.pone.0275631

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*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Desjardins, J., Lofts, A., Campopiano, A., Kennedy, M., Collins, T., Stephenson, S., & Cichonski, M. (2019). *BUCANL/Vised-Marks [MATLAB]*. BUCANL. https://github.com/BUCANL/Vised-Marks.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Palgrave Macmillan.

Dienes, Z. (2019). How do I know what my theory predicts? *Advances in Methods and Practices in Psychological Science, 2*(4), 364–377. https://doi.org/10.1177/2515245919876960

Dunn, L. M., & Dunn, L. M. (1998). Peabody picture vocabulary test, third edition. *Journal of Psychoeducational Assessment, 16*, 334–338.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & van der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, 3*, 3, 1393–1396. https://doi.org/10.1109/ICSLP.1996.607874

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review, 37*, 66–108. https://doi.org/10.1016/j.dr.2015.05.002

Evans, J. L., Saffran, J. R., & Robe, -T. K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 52*(2), 321–335. https://doi.org/10.1044/1092-4388(2009/07-0189

Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage.

Fiveash, A., Bedoin, N., Gordon, R. L., & Tillmann, B. (2021). Processing rhythm in speech and music: Shared mechanisms

and implications for developmental speech and language disorders. *Neuropsychology, 35*(8), 771–791. https://doi.org/10.1037/neu0000766
Flaugnacco, E., Lopez, L., Terribili, C., Zoia, S., Buda, S., Tilli, S., Monasta, L., Montico, M., Sila, A., Ronfani, L., & Schön, D. (2014). Rhythm perception and production predict reading abilities in developmental dyslexia. *Frontiers in Human Neuroscience, 8*. https://doi.org/10.3389/fnhum.2014.00392
François, C., Chobert, J., Besson, M., & Schön, D. (2012). Music training for the development of speech segmentation. *Cerebral Cortex, 23*(9), 2038–2043. https://doi.org/10.1093/cercor/bhs180
François, C., & Schön, D. (2011). Musical expertise boosts implicit learning of both musical and linguistic structures. *Cerebral Cortex, 21*(10), 2357–2365. https://doi.org/10.1093/cercor/bhr022
François, C., Tillmann, B., & Schön, D. (2012). Cognitive and methodological considerations on the effects of musical expertise on speech segmentation. *Annals of the New York Academy of Sciences, 1252*(1), 108–115. https://doi.org/10.1111/j.1749-6632.2011.06395.x
Gabay, Y., Thiessen, E. D., & Holt, L. L. (2015). Impaired statistical learning in developmental dyslexia. *Journal of Speech, Language, and Hearing Research, 58*(3), 934–945. https://doi.org/10.1044/2015_JSLHR-L-14-0324
Gilpin, A. R. (1993). Table for conversion of kendall's Tau to spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educational and Psychological Measurement, 53*(1), 87–92. https://doi.org/10.1177/0013164493053001007
Gingras, B., Honing, H., Peretz, I., Trainor, L. J., & Fisher, S. E. (2015). Defining the biological bases of individual differences in musicality. *Philosophical Transactions of the Royal Society B: Biological Sciences, 370*(1664), Article 20140092. https://doi.org/10.1098/rstb.2014.0092
Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience, 15*(4), 511.
GoldWave Inc. (2022). GoldWave *(6.61)* [Computer software] https://goldwave.com/.
Gu, X., Hoijtink, H., Mulder, J., Lissa, & van, C. J. (2021). *Bain: Bayes factors for informative hypotheses. R package version 0.2.8*. https://CRAN.R-project.org/package=bain.
Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods, 19*(4), 511.
Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximate adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology, 71*, 229–261. https://doi.org/10.1111/bmsp.12110
Harrison, P. M. C., & Müllensiefen, D. (2018a). *Computerised adaptive beat alignment test (CA-BAT), psychTestR implementation*. https://doi.org/10.5281/zenodo.1415353
Harrison, P. M. C., & Müllensiefen, D. (2018b). Development and validation of the computerised adaptive beat alignment test (CA-BAT). *Scientific Reports, 8*(1), Article 12395. https://doi.org/10.1038/s41598-018-30318-8
Henry, M. J., & Herrmann, B. (2014). Low-frequency neural oscillations support dynamic attending in temporal context. *Timing & Time Perception, 2*(1), 62–86. https://doi.org/10.1163/22134468-00002011
Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402. https://doi.org/10.1038/nrn2113
Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods, 24*(5), 539–556. https://doi.org/10.1037/met0000201
Huss, M., Verney, J. P., Fosker, T., Mead, N., & Goswami, U. (2011). Music, rhythm, rise time perception and developmental dyslexia: Perception of musical meter predicts reading and phonology. *Cortex, 47*(6), 674–689. https://doi.org/10.1016/j.cortex.2010.07.010
JASP Team. (2023). *JASP* (Version 0.17.3) [Computer software].
Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.
Kerkhoff, A., Bree, E. D., Klerk, M. D., & Wijnen, F. (2013). Non-adjacent dependency learning in infants at familial risk of dyslexia. *Journal of Child Language, 40*(1), 11–28. https://doi.org/10.1017/S0305000912000098
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. https://doi.org/10.1016/j.neulet.2009.06.030
Kraus, N., Slater, J., Thompson, E. C., Hornickel, J., Strait, D. L., Nicol, T., & White-Schwoch, T. (2014). Auditory learning through active engagement with sound: Biological impact of community music lessons in at-risk children. *Frontiers in Neuroscience, 8*, 351. https://doi.org/10.3389/fnins.2014.00351
Ladányi, E., Persici, V., Fiveash, A., Tillmann, B., & Gordon, R. L. (2020). Is atypical rhythm a risk factor for developmental speech and language disorders? *WIREs Cognitive Science, 11*(5), Article e1528. https://doi.org/10.1002/wcs.1528
Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2017). Statistical learning in specific language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research, 60*(12), 3474–3486. https://doi.org/10.1044/2017_JSLHR-L-16-0439
Langus, A., Boll-Avetisyan, N., Ommen, S., & Nazzi, T. (2023). Music and language in the crib: Early cross-domain effects of experience on categorical perception of prominence in spoken language. *Developmental Science*. https://doi.org/10.1111/desc.13383
Liberto, G. M. D., Pelofi, C., Shamma, S., & Cheveigné, A. de (2020). Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening. *Acoustical Science and Technology, 41*(1), 361–364. https://doi.org/10.1250/ast.41.361
Lizcano-Cortés, F., Gómez-Varela, I., Mares, C., Wallisch, P., Orpella, J., Poeppel, D., Ripollés, P., & Assaneo, M. F. (2022). Speech-to-Speech synchronization protocol to classify human participants as high or low auditory-motor synchronizers. *STAR Protocols, 3*(2), Article 101248. https://doi.org/10.1016/j.xpro.2022.101248
Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Front. Hum. Neurosci., 8*, 213. https://doi.org/10.3389/fnhum.2014.00213
Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *Plos One, 9*(2), Article e89642. https://doi.org/10.1371/journal.pone.0089642
Mandikal Vasuki, P. R., Sharma, M., Ibrahim, R., & Arciuli, J. (2017). Statistical learning and auditory processing in children with music training: An ERP study. *Clinical Neurophysiology, 128*(7), 1270–1281. https://doi.org/10.1016/j.clinph.2017.04.010
Menn, K. H., Ward, E. K., Braukmann, R., van den Boomen, C., Buitelaar, J., Hunnius, S., & Snijders, T. M. (2022). Neural tracking in infancy predicts language development in children with and without family history of autism. *Neurobiology of Language, 3*(3), 495–514. https://doi.org/10.1162/nol_a_00074
Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning, 62*(1), 302–331. https://doi.org/10.1111/j.1467-9922.2010.00626.x
Misyak, J. B., Christiansen, M., & Tomblin, J. B. (2010). On-line individual differences in statistical learning predict language

processing. *Frontiers in Psychology, 1*, 31. https://doi.org/10.3389/fpsyg.2010.00031
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*, Article 101154. https://doi.org/10.1016/j.dcn.2022.101154
Newman, R. S., Rowe, M. L., & Ratner, N. B. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language, 43*(5), 1158–1173. https://doi.org/10.1017/S0305000915000446
Nitin, R., Gustavson, D. E., Aaron, A. S., Boorom, O. A., Bush, C. T., Wiens, N., Vaughan, C., Persici, V., Blain, S. D., Soman, U., Hambrick, D. Z., Camarata, S. M., McAuley, J. D., & Gordon, R. L. (2023). Exploring individual differences in musical rhythm and grammar skills in school-aged children with typically developing language. *Scientific Reports, 13*(1), 2201. https://doi.org/10.1038/s41598-022-21902-0
Ordin, M., Polyanskaya, L., & Samuel, A. G. (2021). An evolutionary account of intermodality differences in statistical learning. *Annals of the New York Academy of Sciences, 1486*(1), 76–89. https://doi.org/10.1111/nyas.14502
Ostrosky-Solís, F., & Lozano, A. (2006). Digit span: Effect of education and culture. *International Journal of Psychology, 41*(5), 333–341. https://doi.org/10.1080/00207590500345724
Palmer, S. D., & Mattys, S. L. (2016). Speech segmentation by statistical learning is supported by domain-general processes within working memory. *Quarterly Journal of Experimental Psychology, 69*(12), 2390–2401. https://doi.org/10.1080/17470218.2015.1112825
Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology, 3*, 320. https://doi.org/10.3389/fpsyg.2012.00320
Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences, 10*(5), 233–238. https://doi.org/10.1016/j.tics.2006.03.006
Pinto, D., Prior, A., & Zion Golumbic, E. (2022). Assessing the sensitivity of EEG-based frequency-tagging as a metric for statistical learning. *Neurobiology of Language*, 1–21. https://doi.org/10.1162/nol_a_00061
Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience, 21*(6), 322–334. https://doi.org/10.1038/s41583-020-0304-4
R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
Rodríguez-Fornells, A., Cunillera, T., Mestres-Missé, A., & de Diego-Balaguer, R. (2009). Neurophysiological mechanisms involved in language learning in adults. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1536), 3711–3735. https://doi.org/10.1098/rstb.2009.0130
Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4
Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science, 12*(4), 110–114. https://doi.org/10.1111/1467-8721.01243
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926
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.
Sallat, S., & Jentschke, S. (2015). Music perception influences language acquisition: Melodic and rhythmic-melodic perception in children with specific language impairment. *Behavioural Neurology, 2015*, Article e606470. https://doi.org/10.1155/2015/606470
Schön, D., & François, C. (2011). Musical expertise and statistical learning of musical and linguistic structures. *Frontiers in Psychology, 2*. https://doi.org/10.3389/fpsyg.2011.00167
Schlichting, L. (2005). *Peabody picture vocabulary test-III-NL*. Hartcourt Assessment BV. https://www.pearsonclinical.nl/ppvt-iii-nl-peabody-picture-vocabulary-test.
Schmalz, X., Altoè, G., & Mulatti, C. (2017). Statistical learning and dyslexia: A systematic review. *Annals of Dyslexia, 67*(2), 147–162. https://doi.org/10.1007/s11881-016-0136-0
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
Shook, A., Marian, V., Bartolotti, J., & Schroeder, S. R. (2013). Musical experience influences statistical learning of a novel language. *The American Journal of Psychology, 126*(1), 95–104.
Siegelman, N. (2020). Statistical learning abilities and their relation to language. *Language and Linguistics Compass, 14*(3), Article e12365. https://doi.org/10.1111/lnc3.12365
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
Silvey, C., Dienes, Z., & Wonnacott, E. (2024). Bayes factors for logistic (mixed-effect) models. *Psychological Methods*. https://doi.org/10.1037/met0000714
Singh, L., Reznick, J. S., & Xuehua, L. (2012). Infant word segmentation and childhood vocabulary development: A longitudinal analysis. *Developmental Science, 15*(4), 482–495. https://doi.org/10.1111/j.1467-7687.2012.01141.x
Smalle, E. H. M., Daikoku, T., Szmalec, A., Duyck, W., & Möttönen, R. (2022). Unlocking adults' implicit statistical learning by cognitive depletion. *Proceedings of the National Academy of Sciences, 119*(2). https://doi.org/10.1073/pnas.2026011119
The MathWorks Inc. (2019). *MATLAB version: 9.6.0.1072779 (R2019a)*. Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com.
Tierney, A., & Kraus, N. (2015). Neural entrainment to the rhythmic structure of music. *Journal of Cognitive Neuroscience, 27*(2), 400–408. https://doi.org/10.1162/jocn_a_00704
Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*(4), 552–564. https://doi.org/10.1037/0096-3445.134.4.552
Van der Wulp, I. M. (2021). *Word segmentation: TP or OCP? A re-analysis of Batterink & Paller (2017)*. Utrecht University. Master Thesis https://studenttheses.uu.nl/handle/20.500.12932/39151.
van der Wulp, I. M., Struiksma, M. E., & Wijnen, F. (2025). Words and meters: Neural evidence for a connection between individual differences in statistical learning and rhythmic ability in infancy. *PsyArXiv*. https://doi.org/10.31234/osf.io/tfusd_v1

Van der Wulp, I. M., Wijnen, F. N. K., & Struiksma, M. E. (2022). *Statistical learning of a new pilot language (Preregistration)*. https://doi.org/10.17605/OSF.IO/WFDKR

Vandermosten, M., Wouters, J., Ghesquière, P., & Golestani, N. (2019). Statistical learning of speech sounds in dyslexic and typical reading children. *Scientific Studies of Reading, 23*(1), 116–127. https://doi.org/10.1080/10888438.2018.1473404

Wechsler, D. (2008). In (4th ed.,, *Vol. 22. Wechsler adult intelligence scale* (pp. 816–827). *San Antonio, TX: NCS Pearson*, 498.

Wang, H. S., Rosenbaum, S., Baker, S., Lauzon, C., Batterink, L. J., & Köhler, S. (2023). Dentate gyrus integrity is necessary for behavioral pattern separation but not statistical learning. *Journal of Cognitive Neuroscience*, 1–18. https://doi.org/10.1162/jocn_a_01981

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Witteloostuijn, M. van, Boersma, P., Wijnen, F., & Rispens, J. (2019). Statistical learning abilities of children with dyslexia across three experimental paradigms. *Plos One, 14*(8), Article e0220041. https://doi.org/10.1371/journal.pone.0220041

Zentner, M., & Strauss, H. (2017). Assessing musical ability quickly and objectively: Development and validation of the Short-PROMS and the Mini-PROMS. *Annals of the New York Academy of Sciences, 1400*(1), 33–45. https://doi.org/10.1111/nyas.13410

Zhang, M., Riecke, L., & Bonte, M. (2021). Neurophysiological tracking of speech-structure learning in typical and dyslexic readers. *Neuropsychologia, 158*, Article 107889. https://doi.org/10.1016/j.neuropsychologia.2021.107889

Zhang, Z., & Wang, L. (2017). Advanced statistics using R. *Granger, IN: ISDSA Press*. https://advstats.psychstat.org. ISBN: 978-1-946728-01-2.

Zuk, J., Vanderauwera, J., Turesky, T., Yu, X., & Gaab, N. (2022). Neurobiological predispositions for musicality: White matter in infancy predicts school-age music aptitude. *Developmental Science*., Article e13365. https://doi.org/10.1111/desc.13365