Investigating individual differences in linguistic statistical learning and their relation to rhythmic and cognitive abilities

A speech segmentation experiment with online neural tracking

I. M. (Iris) van der Wulp[a,*]
M. E. (Marijn) Struiksma[a]
L.J. (Laura) Batterink[b]
F. N. K. (Frank) 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

[*]*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

Acknowledgements

We would like to sincerely 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 ~~an~~ earlier versions of this manuscript.
This work is funded by the Netherlands Organization for Scientific Research (NWO), project number PGW.21.007.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

Statistical learning, speech segmentation, individual differences, neural oscillations, EEG, phase-locking, rhythmic abilities, cognitive abilities.

Abstract

Objective: Statistical Learning (SL) is an essential mechanism for speech segmentation. Importantly, 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 is to contribute to uncovering the underpinnings of such individual differences in auditory SL for word segmentation. We hypothesize that individuals with better musical – specifically rhythmic – abilities will show better SL for speech segmentation.

Methodology: Participants will be exposed to an artificial language consisting of trisyllabic nonsense words. Recent methodological innovations allow online assessment of SL via *electroencephalography* (EEG) measures of neural entrainment. The current study will use this EEG method to measure individual SL performance during exposure. ~~Moreover, we will also assess learning post-exposure using behavioral tasks of explicit and implicit memory.~~ Aiming to assess individual differences, we will link the neural measure~~s~~ of SL to a battery of tests assessing possible individual differences by measuring rhythmic, musical, and cognitive abilities, as well as vocabulary size.

Expected results: We predict that individuals with better rhythmic abilities will show greater neural entrainment to external auditory rhythms, supporting better extraction of the transitional probabilities between syllables. Specifically, we expect to see greater neural entrainment in these individuals to the frequency of the tri-syllabic words in our stimuli, indicative of SL, than individuals with lower scores on the rhythm perception tasks. ~~We also anticipate behavioral evidence of better SL performance in individuals with rhythmic abilities.~~ Furthermore, we ~~predict that~~explorativeIy investigate if larger working memory capacity contributes to better SL as captured online by the EEG measure. The question of whether vocabulary size in adulthood contributes to better SL is also explorative, as the connection between SL and vocabulary size has predominantly been researched in children. If this association persists in the adult population, it is anticipated to manifest as a positive correlation.

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 are subsequently 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 is 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

[1] This is also frequently referred to as *word segmentation.*
[2] Syllables are a basic unit of spoken language (e.g., Assaneo & Poeppel, 2020) and therefore transitional probability computations are made based on neighboring syllables for speech segmentation.

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 (François, Tillmann et al., 2012; Bogaerts et al., 2022). 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 (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 towards the word-final syllables, which are the most predictable compared to the second and first syllable.

Implicit measures such as the target detection task 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., 2022). 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 ms) and the tri-syllabic words (1.1 Hz; 900 ms). 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 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 target detection task. 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., 2022), 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 et al., 2010; Misyak & Christiansen, 2012; Siegelman & Frost, 2015; Siegelman, 2020).

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

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: <u>what underlies individual differences in SL, which in turn might predict inter-individual variation in language attainment?</u> 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 & 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; Kaufman et al., 2010; 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 both a behavioral and neural 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 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 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

[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*.

[5] Happening at regular intervals. In this case, all syllables were 111 ms long, creating a constant syllable frequency of 4.5 Hz (see Assaneo et al., 2019, p. 7).

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 whose 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 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, Chobert and colleagues (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, Tillmann, and colleagues (2012) hypothesize 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 and 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).

[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.

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 (Assaneo & Poeppel, 2020). This converges with the hypothesis by François, Tillmann 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 and colleagues (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 beforementioned 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 (Cirelli et al., 2016; Daikoku & Goswami, 2022; Fiveash et al., 2021; Giraud & Poeppel, 2012; Liberto et al., 2020; Menn et al., 2022; Nozaradan et al., 2011; Peelle & Davis, 2012; Poeppel & Assaneo, 2020; Tierney & Kraus, 2015). Furthermore, *rhythmic ability* – the ability to accurately detect and (behaviorally) synchronize to an auditory pulse – has been found to

[7] Regular patterns of strong and weak beats.

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 will investigate 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 will be 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 will be an extension of prior work in multiple ways. In order to investigate individual differences, we will measure participants' performance on tasks assessing musical, rhythmic, linguistic, and general cognitive abilities. We will then relate these scores to the neural measure of SL. To our knowledge, a relation between musical/rhythmic abilities and SL

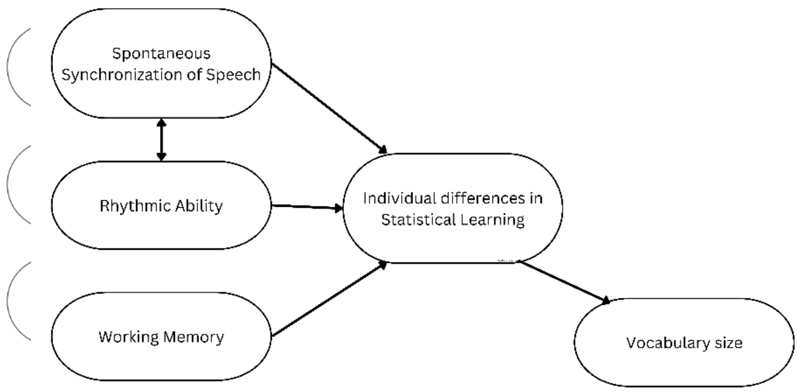


Figure 1. Predictions of the current study represented graphically.

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 Figure 1 and the paragraphs below for our predictions regarding the individual differences and SL.

We predict that rhythmic ~~and musical~~ abilities positively correlate with SL performance. We will test rhythm perception using two tasks (Harrison & Müllensiefen, 2018a, 2018b; Zentner & Strauss, 2017) ~~in addition to a questionnaire about general musical ability and musical training experience (Bouwer et al., 2016; Müllensiefen et al., 2014).~~ We predict these tasks to be ~~inter-~~positively correlated, but we use multiple tasks to be sure that we measure rhythm perception as accurately as possible. We will also measure behavioral rhythmic speech-to-speech entrainment by using the SSS task (Assaneo et al., 2019). We expect performance on this task to also be a ~~significant~~ predictor of SL, which would replicate a key finding reported by Assaneo and colleagues (2019). We will perform a mediation analysis to further investigate interrelations between these rhythm tasks, the SSS task, and SL ability (see section 2.6 for details). I~~i~~n addition, we exploratively add ~~to~~ a questionnaire about general musical ability and musical training experience (Bouwer et al., 2016; Müllensiefen et al., 2014).

Moreover, we will broaden 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 digit span is associated with verbal working memory and depends on the phonological loop, which is the most interesting for our study. The backward digit span, however, is more so 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 will exploratively investigate whether working memory aids SL online.

In addition, we will administer 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 – proposed in 2.3.3.*d* – in that it required participants to choose a synonym for a target word, whereas our proposed vocabulary test requires participants to choose a picture corresponding to the meaning of a target word. Therefore, analogous to earlier research with children, we predict a positive relation between SL and vocabulary size.

Finally, even though this experiment will answer the new questions above, it will also be a partial replication and extension of earlier experiments (Assaneo et al., 2019; Batterink & Paller, 2017; Pinto et al., 2022). We therefore expect 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 predict 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). Moreover, we will test if the neural measure of SL correlates ~~significantly~~ positively with the behavioral tasks (Batterink & Paller, 2017). We are extending this prior work because the participants in the current study will be speakers of Dutch, and the stimuli we have are newly created and adhere to Dutch phonotactics.[8] Finally, we expect to replicate ~~effect of the SSS task showing~~ 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).

2. Materials and methods

2.1. Participants

We will start with an initial sample of 45 participants with data useable for analysis, identical to Batterink and Paller (2017). Then, we will perform 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 is reached for our *critical analyses*, or when we reach a maximum feasible sample of 105 participants. We performed simulations[9] on our proposed statistical models (see sections 2.4-2.6) and also simulation-based Bayes Factor Design Analysis (BFDA; Schönbrodt & Wagenmakers, 2018; Schönbrodt & Stefan, 2019) for simulations of correlations. Details on these simulations can be found in appendix A and the supplementary materials. We chose 15 participants as the updating sample size, because this reflects approximately two to three weeks of data collection. We will then use a third or fourth week to re-run the analyses and to determine if we need to add another sample. This way, we can create a monthly updating cycle. The critical analyses (marked green in the study design table in appendix A) are the following:

[8] More details on the methodology used to create these stimuli are described in van der Wulp et al. (2022). See also appendix B for details on a pilot experiment with these stimuli.
[9] Link to our simulations supplement: https://osf.io/jhbe8/files/osfstorage/6568489a56f9cf04a440a7e1

- The analysis for the replication of the EEG results of Batterink & Paller (2017; see section 2.4.1), with regard to a difference in the WLI between the structured and random conditions.
- The correlations between the tests for rhythmic ability; PROMS, CA-BAT and SSS (see section 2.6), in order to be able to perform the mediation analysis.
- Evidence for or against a direct effect of SSS PLV on the WLI, in order to be able to perform the mediation analysis (see section 2.6).
- ~~Correlations calculated for the WLI with vocabulary and working memory if they are not added to the mediation (see section 2.6).~~

Participants will not be invited to participate if they report 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 can be excluded after participation in the case of technical issues that cause a premature termination of the experiment, if the participant wishes to retract/stop their participation during the experiment, or if the participant has < 50% targets detected in the target detection task. In our pilot experiment (appendix B) and earlier studies from Laura Batterink, all participants performed above this percentage.

Participants will all be native speakers of Dutch and they will be between 18 and 35 years old. The experiment is approved by the Linguistics Chamber of the Faculty Ethics Assessment Committee of Humanities at Utrecht University (reference number: LK-22-174-02), and participants will be compensated with a €20 gift card for their time (the session will take approximately two hours).

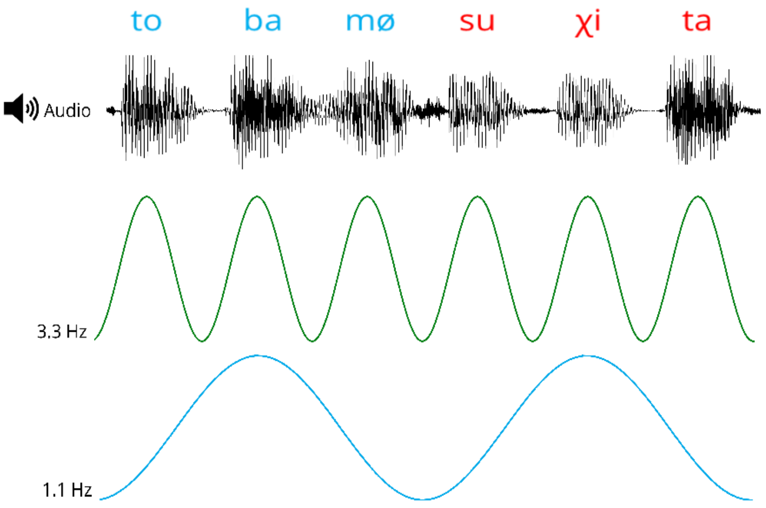


Figure 2. Stimuli and stimulus frequencies in the structured stream. The audio represents the depicted syllables. The syllables of the same color form a word. The green waveform depicts the syllable frequency of 3.3 Hz. The blue waveform depicts the tri-syllabic word frequency of 1.1 Hz.

2.2. Stimuli

The stimuli consist of syllables which are combined into tri-syllabic nonwords (from now on referred to as 'words') that adhere to Dutch phonotactics and have been piloted for their learnability (see appendix B for details on the pilot experiment). The syllable inventory consists of 12 syllables, from which four words are formed for the *structured condition*: /suχita, tobamø, sytøbo, χøbyti/. In the structured stream, the transitional probabilities of neighboring syllables are 1.0 within a word and 0.33 between words. The word order is pseudorandomized, such that the same word does 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 is concatenated in a pseudorandom order, under the constraint that the same syllable cannot consecutively repeat (as in Batterink & Paller, 2017). This yields a transitional probability of 0.09 throughout the random condition. The syllables used in this condition are: /da, pø, nu, dø, χo, py, ro, dy, sa, χy, ri, sø/, corresponding to set *B* in the pilot experiment (see appendix B and C: 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 are 300 ms long (100 ms consonant, 200 ms vowel), creating a word-length of 900 ms. Thus, this yields a syllable frequency of 3.3 Hz and a word or triplet frequency of 1.1 Hz (see Figure 2). We generated coarticulated speech streams of 13.5 minutes per condition in total, divided over three blocks of 4.5 minutes. Each block is 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 seconds at the beginning and end of each block, to avoid a segmentation cue at the beginning of the stream. Stimuli will be 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 can be read as EEG markers with Presentation. The EEG markers and their corresponding syllables can be found in table C1 in appendix C.

2.3. Procedure

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

2.3.1. *Listening Task*

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

Participants will first perform the listening task in the structured condition. After this, the rating task and target detection task (see 2.3.2.) will be administered, followed by another iteration of the listening task to the random stream. The listening task will be divided into three blocks of 4.5 minutes per condition, yielding 13.5 minutes per condition and 27 minutes in total for both conditions. Participants will take a short break between blocks.

2.3.2. *Behavioral tasks of SL outcomes*

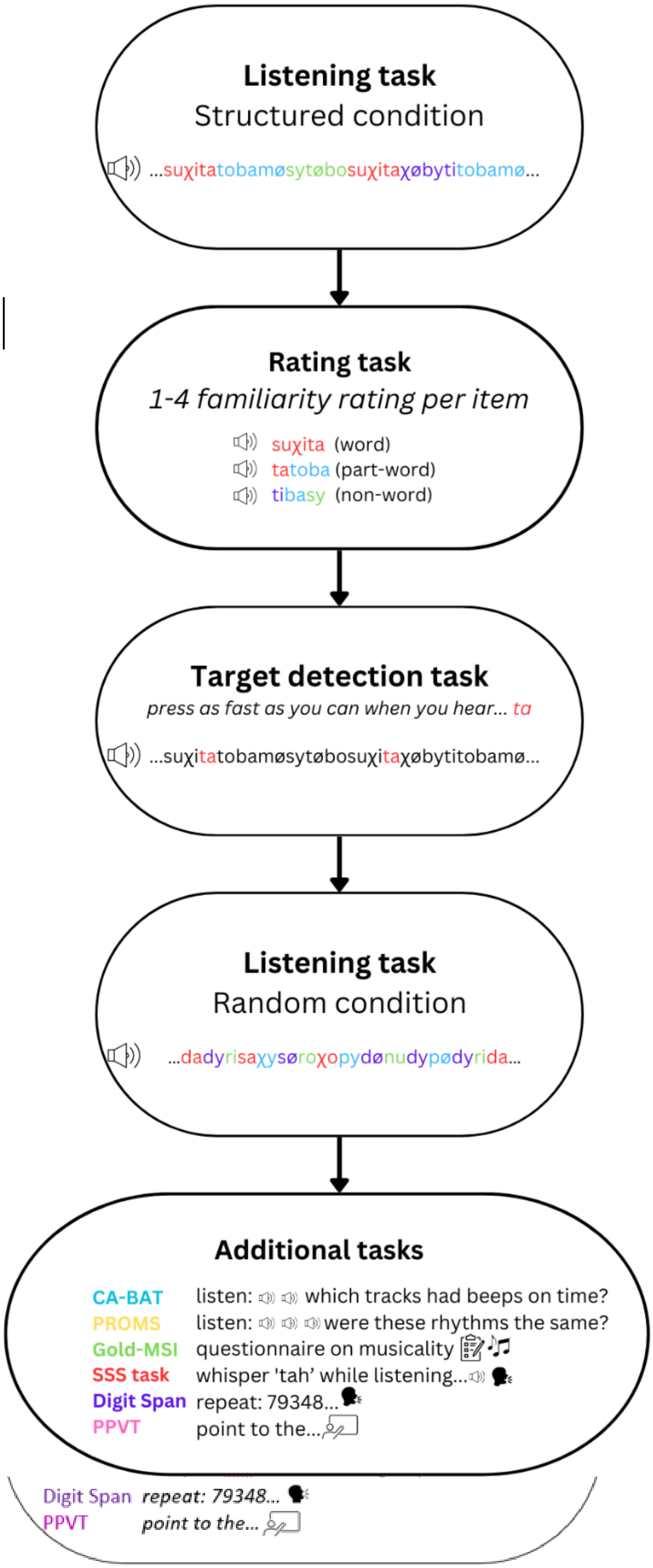


Figure 3. Schematic overview of the experimental procedure.

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

With respect to the rating task, participants will be auditorily presented with a word or foil in each trial. The foils can 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 Figure 3; see table C2 in appendix C for the full list of foils). There will be 16 trials consisting of the four words from the listening task, all eight possible part-words and four non-words. On each trial, participants will rate on a four-point scale how familiar the word is to them (scale: unfamiliar – fairly unfamiliar – fairly familiar – familiar).

The second post-learning task our participants will perform is the target detection task (Batterink, 2017; Batterink et al., 2015; Batterink & Paller, 2017, 2019). Participants will be presented (auditorily and visually) with a target syllable and subsequently hear 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 are asked to press a button as quickly and accurately as possible when they hear the target syllable. For each target syllable there are 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 will employ three measures assessing rhythmic and musical abilities of the participants. First, participants will perform the Computerized Adaptive Beat Alignment Test (CA-BAT; Harrison & Müllensiefen, 2018a, 2018b), in which participants listen to the same piece of music twice, accompanied by beeps in two conditions. In one condition, the beeps are synchronized with the rhythm of the music, and in the other condition, the beeps are not synchronized with the rhythm of the music. Participants indicate which of the two tracks had the beeps in sync with the rhythm of the music.

Second, participants will complete 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 listen twice to the same rhythm and then to a third rhythm. Participants then indicate whether the third rhythm was identical or different compared to the first two.

Third, participants will complete 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 consists of the following sub-scales: active engagement with music, perceptual abilities, musical training, singing abilities and emotional engagement.

b. Spontaneous Synchronization to Speech

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

c. Working memory

Participants will perform a forward digit span (Wechsler, 2008) as an indication of working memory capacity. In this test, the experimenter orally names digits and the participant is instructed to repeat them. The number of digits will increase until the participant fails to remember one or more digits in the array.

d. Vocabulary

Finally, we will administer 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 will be 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 will be made with the Active-Two system (Biosemi, Amsterdam, The Netherlands). Additional electrodes will be placed on the left and right mastoid, above and below the left eye, and at the outer canthi of both eyes. Scalp signals will be 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 will be kept below 20 mV. If the impedance of a channel is higher than this, it will be labeled as a bad channel during data collection to be interpolated during data analysis.

The EEG data will be analyzed in MATLAB (The MathWorks Inc., 2019) using EEGLAB (Delorme & Makeig, 2004) and the ERPLAB open-source toolbox (Lopez-Calderon & Luck, 2014). The data will be bandpass filtered from 0.1 to 30 Hz and 50 Hz notch filtered offline. Bad channels identified upon visual inspection of the data or during data collection will be interpolated. Data sections comprising large artifacts will also be identified through visual inspection and manually rejected. A channel is labeled as bad during the analysis if it was labeled bad during data collection due to high impedance, or if it shows frequent noise or drifts upon visual inspection of the data. Eye movement artifacts will be 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). In case of excessive artifacts for a given participant, we will use Independent Component Analysis (ICA) to remove only the artifactual components from the data (Moreau et al., 2022). Finally, data of participants that do not show a clear ITC peak at the syllable frequency of 3.3 Hz, indexing basic auditory processing of the syllables, will be excluded.

We will time-lock the data to the onsets of the tri-syllabic words and divide it into non-overlapping epochs of 10.8 seconds, corresponding to the duration of 12 trisyllabic words (36 syllables). We will then quantify 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 will be calculated after a Fast Fourier Transform (FFT) for each epoch across frequency bins of interest: between 0.6 to 5 Hz, with a bin width of 0.09 Hz (following Batterink & Choi, 2021; Benjamin et al., 2021; Moreau et al., 2022). The Word Learning Index (WLI) will then be 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 will follow the methodology of Moreau et al. (2022) using a sliding window to map learning trajectories during the listening task. We will create *epoch bundles* each containing 5 epochs, with each bundle shifted by one epoch (e.g., epochs 1-5, 2-6, 3-7, etc.). This will result in 54 seconds of exposure per bundle. We will compute this for the 20 fronto-central electrodes previously used by Moreau et al. (2022)[11].

2.4.1. *Statistical analyses of the neural data*

We will statistically test for ~~significance~~ 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 Schmalz et al., 2023; Dienes, 2019). We will calculate the ITC for the word and syllable frequencies over the exposure period and use them to compute the WLI, as described in 2.4. above. We will then conduct 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 will have the WLI as the dependent variable and we will include a random slope for language condition (structured/random) per participant. We expect 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,

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

2017; Moreau et al., 2022; Pinto et al., 2022; van der Wulp, 2021). We will statistically determine this by including condition as a predicting factor. If we find evidence for an effect of condition, we will test for an interaction of condition and epoch bundle number as the predicting factors.

We will then compute the BF following Silvey et al. (2022). We specify our model of H1 for the condition effect as a half-normal distribution with a mean of 0 and an SD of 0.19~~ / 2 = 0.095~~, corresponding to the estimate for the original condition effect of Batterink & Paller (2017). For the interaction effect, we will follow the same procedure while our SD is 0.07~~ / 2 = 0.035~~. See the simulation supplement for the models yielding these estimates on the data of Batterink & Paller (2017). If we encounter singularity errors~~,~~ or if the model does not converge, we will first remove the correlations between random slopes. If it still does not converge or still is singular, we will remove the random slope. If the model does not converge, we will collect another sample of 15 participants (see sampling plan in 2.1)~~. If~~ until we have reached our maximum sample size~~, we will simplify the model by removing the random slope~~.~~. If that does not yield reliable results, we will remove the interaction and test for the condition effect alone.~~

We will follow the analyses with sensitivity analyses reporting a robustness region (Dienes, 2019). We will test for prior models of H1 where the condition effect is 0, to 0.38 for the condition effect (twice as large as the effect found by Batterink and Paller, 2017) to find the region where the BF_{10} is still > 3 or < 1/3. We choose 0.38 as the maximum, because in theory the WLI can range until infinity, and we do not expect the effect to be more than twice as large. In similar vein, we will test for robustness of the interaction between 0 and 0.14.

2.5. Behavioral data analyses

2.5.1. *Group Analyses of behavioral SL outcome measures*

The dependent variable for the rating task consists of the familiarity ratings on the four-point scale. Random effects will be random intercepts for participant and item. We will test 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 will use 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:

$$h(\boldsymbol{\theta}) = N(\mathbf{0}, \textstyle\sum_{\infty}) \quad (1)$$

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:

$$g(\boldsymbol{\theta}|\boldsymbol{X}) \approx N(\hat{\boldsymbol{\theta}}, \textstyle\sum_{\boldsymbol{\theta}}) \quad (2)$$

Where $\boldsymbol{X}$ 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 H_c as the ratio of the posterior and prior probabilities that the inequality constraints hold:

$$BF_{ic} = \frac{f_i}{c_i} \times \frac{1-c_i}{1-f_i} \quad (3)$$

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 formulate the following informative hypothesis for Bain, which will be evaluated against its complement (Equation 3):

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

After the initial analysis, we will also conduct 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).

~~After the initial analysis in Bain, we will also conduct a sensitivity analysis as described in Hoijtink et al. (2019).~~

With respect to the target detection task, RTs are only taken into consideration for any of the analyses if the button press occurred within 1200 ms 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 will be 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 will furthermore add a random intercept for participant to account for individual differences in baseline RTs. Finally, we will add the variable stream position as a covariate, referring to the trial number of the target syllable in the stream, 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 will use 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~~ / 2 = 15.96~~, which was the result of our pilot experiment on the target detection task (see appendix B).

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

2.5.2. *Correlations between neural and behavioral SL data*

For the rating task, we will compute a composite *rating score* for each participant, following Moreau et al. (2022; Batterink & Paller. 2017, 2019), subtracting the mean rating for foils (part-words and non-words) from the mean rating score for words. For the target detection task, we will calculate a *RT facilitation score* for each participant (Batterink & Paller 2019; Moreau et al., 2022), by subtracting the RTs for the third syllable from the RTs for the first syllable and dividing this by the RTs for the first syllable: ($RT\ facilitation = (RT_{S1} - RT_{S3})/RT_{S1}$), which accounts for individual baseline RTs. We will conduct Bayesian correlation analyses between the overall WLI in the structured condition, the rating score, and the RT facilitation score to determine whether individual variability in neural entrainment during exposure is related to subsequent SL performance. We will perform 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 Figure 4). The width is inversely related to the parameters of the beta distribution. For instance, a prior weight of 0.5 generates a beta(2,2) stretched from -1 to 1 (2 = 1/0.5). In this case, the beta distribution is cut in half at 0, because we only hypothesize positive correlations. Since the effects in Batterink and Paller (2017) were *r* = 0.32 for the rating task, and *r* = 0.42 for the TDT, we will ~~, adhering~~ adhere to ~~the uniform default~~the prior κ = 0.5, which places

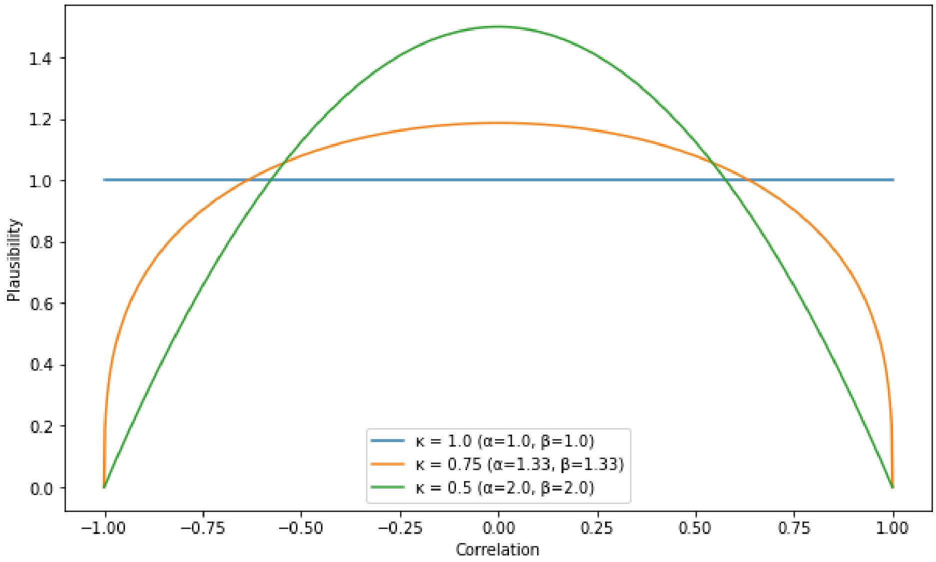


Figure 4. Beta prior distributions in JASP for correlations. In JASP, one specifies the width of the prior distribution (κ). The width is inversely related to the parameters of the beta distribution. The default value of κ is 1 (blue line). We will use κ = 0.5 (green line) for medium and κ = 0.75 (orange line) for large hypothesized correlations. When testing one-sided, the distribution is cut in half at 0.

less prior weight on big effect sizes and relatively more around 0. We will follow 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).~~1.~~

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 0-8) and the accent sub-test (between 0-10), the mean of which we will record 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 1-7 for each participant and sub-scores also ranging between 1-7 per sub-scale.

For the SSS task (Assaneo et al., 2019), we will adhere to the protocol described in Lizcano-Cortés et al. (2022). We will calculate 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 will measure the longest span for each participant. This will then be recorded as one data point per participant.

Finally, for the PPVT-III-NL (Dunn & Dunn, 1998; Schlichting, 2005), raw scores will also be recorded as one data point per participant.

All scores on the individual differences' tests will be standardized before statistical analyses are conducted. This will be 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 will first perform 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 will perform these correlations using the statistical software JASP (JASP Team, 2023). With regard to the priors for these correlations, w~~, adhering to the uniform default prior κ = 1. W~~e expect the measures of rhythm (e.g., CA-BAT, PROMS, and SSS task PLV) to be highly positively ~~-~~correlated. Therefore, we will use the prior κ = 0.75, 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 is also hypothesized to have a positive correlation with the rhythm tasks, but we ~~We~~ do not necessarily expect correlations between the Digit Span, PPVT-III-NL, and rhythm tasks. For these explorative correlations, we will adhere to the prior κ = 0.5, which places less prior weight on big effect sizes and relatively more around 0. This gives us a reasonable chance of finding a theoretically interesting medium-to-large effect size if it exists (see also our simulations

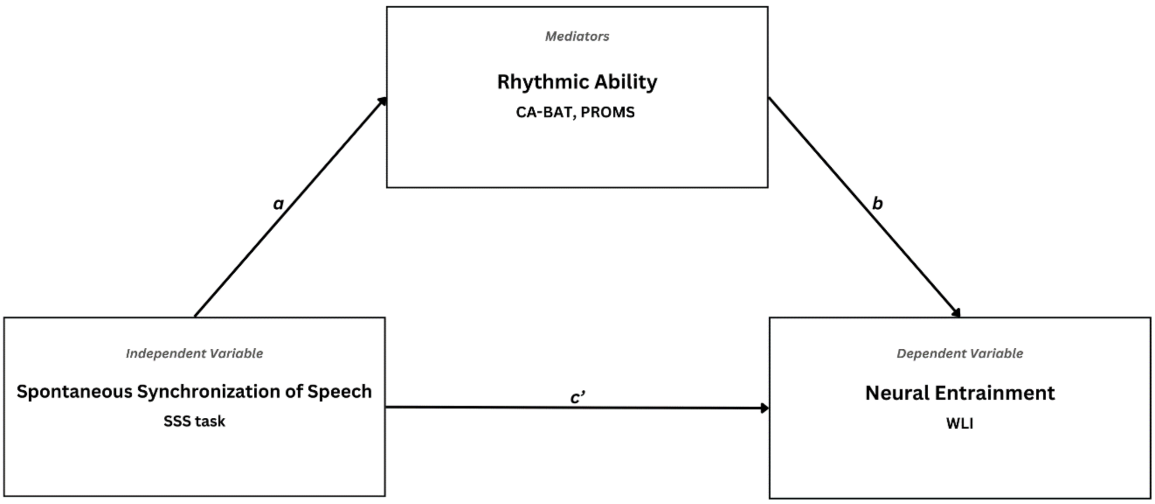
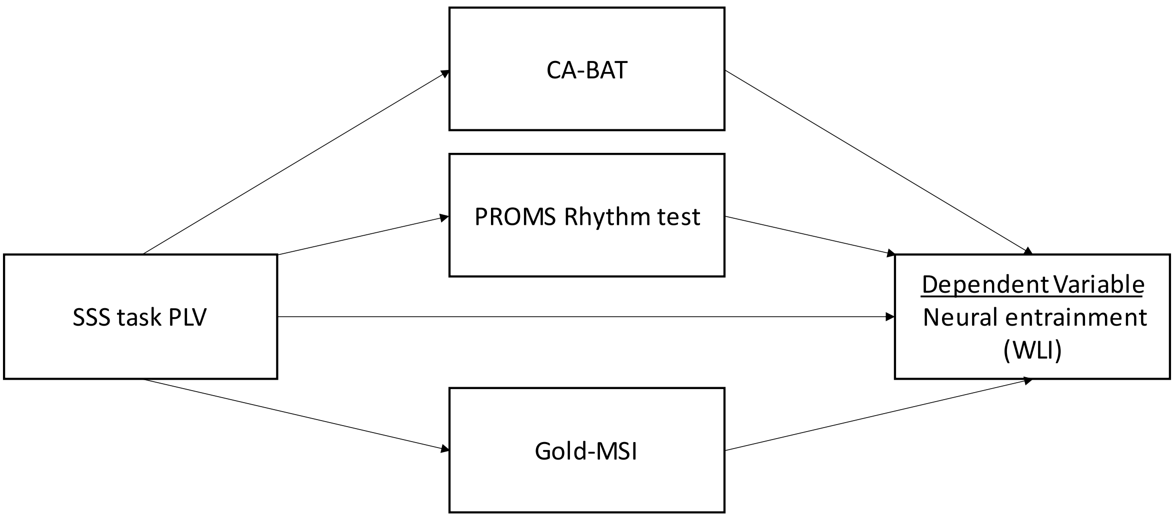


Figure 54. Proposed mediation analysis, hypothesizing a direct effect of SSS PLV (spontaneous synchronization of speech) on the WLI (neural measure of SL) in the structured condition, adding the CA-BAT, PROMS (both rhythmic ability), as mediators. The *c'* path denotes the direct effect, and the path *ab* denotes the mediated effect.

supplement and appendix A). We will follow these analyses with sensitivity analyses provided in JASP (see section 2.5.2).

Subsequently, in order to assess the influence of our predictors for individual differences on SL, we will perform a mediation analysis with multiple mediators (e.g., Dienes, 2019; Field, 2013; Zhang & Wang, 2017). The WLI in the structured condition will be the dependent variable, and we predict 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 will test for this direct effect initially by performing a ~~correlation~~ regression ~~between~~ of the SSS task ~~and~~ on the WLI, ~~using~~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 statistical software JASP (JASP Team, 2023), adhering to the uniform default prior κ = 1~~. The hypothesis for a null effect will be 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, as well as the simulations supplement for code implementation. We hypothesize that ~~this~~ the direct effect, if found, is mediated – and can perhaps be completely explained – by one or more of our measures ~~for~~ of ~~musical and~~ rhythmic ability. Figure 5~~4~~ depicts the planned mediation analysis. We will perform the full mediation analysis using the lavaan package in R (Rosseel, 2012), and will subsequently load the model into Bain (Gu et al., 2021; Hoijtink et al., 2019). We will evaluate the mediators in Bain under the informative hypotheses *a-path* > 0 & *b-path* > 0 (e.g., Miočević et al., 2020). After the analyses in Bain, we will also conduct sensitivity analyses as described in section 2.5.1.



We will, however, only add tasks as mediators that ~~significantly~~ positively correlate~~d~~ with the SSS task ~~in the correlation analysis between all tasks above~~. This could mean that the

Digit Span and/or PPVT-III-NL will be additionally added as mediators, or that one or more of the rhythm tasks is not added. For tasks that do not correlate with the SSS task, we will perform explorative correlations between these tasks and the WLI, using JASP with the prior κ = 0.5 and sensitivity analyses as described above. The scenario outlined ~~above~~ in Figure 5 is created under the hypothesis that the ~~rhythm~~ SSS task~~s~~ does not correlate with the Gold-MSI, Digit Span and the PPVT-III~~.~~. ~~If this is indeed the case, we will perform correlations separately between the WLI in the structured condition and the Digit Span, as well as the PPVT-III, respectively. If there are other tasks that do not correlate with the SSS task and that will thus not be added to the mediation analysis, we will also separately correlate these with the WLI. All these correlations will be performed in JASP (JASP Team, 2023), adhering to the uniform default prior κ = 1. We will perform the full mediation analysis using the lavaan package in R (Rosseel, 2012), and will subsequently load the model into Bain (Gu et al., 2021; Hoijtink et al., 2019), under the informative hypothesis for the direct effect: *c-path* > 0. We will evaluate the mediators in Bain under the informative hypotheses *a-path* > 0 & *b-path* > 0. This approach for a mediation analysis in Bain was previously established by Miočević et al. (2020). After the initial analysis in Bain, we will also conduct a sensitivity analysis as described in Hoijtink et al. (2019).~~

References

Arciuli, J. (2017). The multi-component nature of statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 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

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. 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*, 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 2022.11-16. 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), 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

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*, 1393–1396 vol.3. 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, -Torres Kathryn. (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., 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

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), 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., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*(4), 511.

Gu, X., Mulder, J., and 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. DOI: 10.1111/bmsp.12110

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

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), 12395. https://doi.org/10.1038/s41598-018-30318-8

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 edition). 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

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

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), 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), 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.

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

The MathWorks Inc. (2019). MATLAB version: 9.6.0.1072779 (R2019a), Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com

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

Miočević, M., Klaassen, F., Geuke, G., Moeyaert, M., & Maric, M. (2020). Using Bayesian methods to test mediators of intervention outcomes in single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, *14*(1–2), 52–68. https://doi.org/10.1080/17489539.2020.1732029

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*, 101154. https://doi.org/10.1016/j.dcn.2022.101154

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), e89642. https://doi.org/10.1371/journal.pone.0089642

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

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. R Foundation for Statistical Computing*, Vienna, Austria. 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

Rosseel, Y. (2012). Lavaan: An R package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1-36. URL http://www.jstatsoft.org/v48/i02/

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. (1996). 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. (1996). 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*, e606470. https://doi.org/10.1155/2015/606470

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

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

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schönbrodt, F. D. & Stefan, A. M. (2019). BFDA: An R package for Bayes factor design analysis (version 0.5.0) Retrieved from https://github.com/nicebread/BFDA

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), e12365. https://doi.org/10.1111/lnc3.12365

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. (2021). *Bayes factors for mixed-effects models.* PsyArXiv. https://doi.org/10.31234/osf.io/m4hju

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

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

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

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

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

Wechsler, D. (2008). Wechsler adult intelligence scale—Fourth edition (WAIS-IV). *San Antonio, TX: NCS Pearson*, *22*(498), 816–827.

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

Winter, B. (2020). Statistics for linguists : an introduction using R. Routledge. https://doi.org/10.4324/9781315165547

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), e0220041. https://doi.org/10.1371/journal.pone.0220041

Wulp, I. M. van der (2021). *Word segmentation: TP or OCP? A re-analysis of Batterink & Paller (2017)* [Master Thesis, Utrecht University]. https://studenttheses.uu.nl/handle/20.500.12932/39151

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

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*, 107889. https://doi.org/10.1016/j.neuropsychologia.2021.107889

Zhang, Z. & Wang, L. (2017). *Advanced statistics using R.* [https://advstats.psychstat.org]. Granger, IN: ISDSA Press. 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, e13365*. https://doi.org/10.1111/desc.13365

Appendix A
Study Design Table

NOTES:

- Our ~~alpha level~~inference criterium is a Bayes Factor (*BF_{10}*) > 6 or ~~or BF_{01}~~ < 1/6. If we have reached our maximum sample size (see sampling plan), and we do not reach BF_{10} < 1/6, while having reached BF_{10} < 1/3, we interpret this as moderate evidence for H0.
- Neural entrainment will be expressed by the Inter-Trial Coherence (ITC) and from the ITC to the word (1.1 Hz) and syllable (3.3 Hz) frequencies we will calculate the Word Learning Index (WLI; Batterink & Paller, 2017). See sections 1.2 and 2.4 in the report for more details on the WLI computation.
- We will run our EEG analyses on the 20 fronto-central electrodes previously used by Moreau et al. (2022). See section 2.4 of the report.
- Some of the power simulations are based on a student pilot: this is a MA thesis project conducted in our lab, which yielded data for N = 15 for the tests of individual differences. This data was not analyzed as part of the MA student's project but could be used as input for our power and effect size estimations. For more details about the student project, see: www.doi.org/10.17605/OSF.IO/MA2C6.

Question	Hypothesis	Sampling plan	Analysis Plan	Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes
RQ1a. Can we replicate Batterink & Paller (2017)'s findings that the WLI is higher in the structured than in the random condition?~~,~~ **RQ1b.** ~~and~~ Can we replicate	We hypothesize that we will replicate Batterink & Paller (2017)'s effect of neural entrainment by finding a difference in our independent within-participant variable *language condition*. We expect a higher WLI in the	We will start with an initial sample of 45 participants, replicating Batterink & Paller (2017). Then, we will perform Bayesian updating, by repeating the statistical analyses after every added	We will create a LMM with the following syntax: WLI(per epoch bundle; N of data points per participant vary; standardized) ~ 1 + condition(structured/random) + (1 + condition | participant). ~~WLI(per epoch bundle; N of data points per participant vary; standardized) ~ 1 + condition(structured/random) *~~	Our planned sample size for RQ1 is based on being able to test for the effect of condition (H1a). This includes the following arguments: (1) Batterink and Paller (2017) included a sample of 45 participants and report in a subsequent publication (Choi et al., 2020, p. 1163) an estimated Cohen's *d* effect	Evidence for (as expressed by BF_{10} > 6) H1a ~~A main effect of condition~~ in the expected direction would indicate that we replicated the findings of Batterink & Paller (2017)~~,~~ ~~;~~ indicating stronger relative entrainment to words in the structured conditions compared to the random condition. ~~entailing and that participants learned the words through SL based on~~	The neural entrainment based WLI provides an accurate and sensitive measure of~~to~~ SL for speech segmentation.

Batterink & Paller (2017) in that there is an interaction between this effect of condition and exposure time?	structured condition of the listening task compared to the random condition (**H1a**). Furthermore, we hypothesize a WLI increase over the course of learning in the structured but not the random condition **(H1b)**. This would be attested by an epoch bundle * condition interaction.	sample of 15 participants, until the threshold value of $BF_{10} > 6$ or $BF_{\cancel{0}10} < 1/6$ is reached for our critical analyses, or when we reach a maximum feasible sample of 105 participants. See the clarification of **Bayesian Updating** below the table.	~~epoch bundle$_{(1:N\ of\ epochs)}$ + (1 + condition | participant).~~ If we encounter singularity errors~~,~~ or if the model does not converge, we will first remove the correlations between random slopes. If it still does not converge or still is singular, we will remove the random slope. If the model still does not converge, we will collect another sample of 15 participants (see sampling plan) until we reach our maximum sample size. ~~If we have reached our maximum of 105 participants, we will simplify the model by removing the random slope.~~ To calculate the BF, we will follow Silvey et al. (2022). We specify our model of H1 as a half-normal distribution with a mean of 0 and an SD of 0.19, corresponding to the estimate for the original condition effect of Batterink & Paller (2017).	size of 0.56 and power of .98 for the WLI difference between the structured and random conditions in their 2017 study. (2) Furthermore, the data of Batterink & Paller (2017) have also been reanalyzed using a Linear Mixed Modelling analysis approach (van der Wulp, 2021, p. 24), yielding similar results as the original. (3) Moreau et al., (2022) found ~~a significant~~ an increasing WLI per epoch bundle in their adult sample (N =24). They only presented a structured condition. See table 1 in their publication (p. 6). (4) We also simulated data based on the WLI values of Batterink & Paller (2017). See the simulation results under **RQ1** below this table.	~~TPs in the structured condition, but not in the random condition.~~ ~~No main~~Evidence for H0 (expressed by $BF_{10} < 1/6$) showing that there is no effect of condition would indicate that we have ~~no~~ evidence indicating ~~that participants did not acquired the words~~similar entrainment to words versus syllables between the structured and random conditions. The time course analysis (H1b) could shed more light on the origin of such a result if this is the case, as could the behavioral tests of learning (RQ2). Evidence for a~~A~~n interaction between condition and epoch bundle in the predicted direction would indicate that we have further replicated the findings of Batterink & Paller (2017) by showing a progressive learning trajectory in the structured,	

				We will follow this analysis with a sensitivity analysis reporting a robustness region (Dienes, 2019). We will test for prior models of H1 where the condition effect is 0, to 0.38 (twice as large as the effect found by Batterink and Paller, 2017) to find the region where the BF_{10} is still > 3 or < 1/3. See section 2.4.1. for more information. If the model converges and provides evidence for H1 for the condition effect, we will test for H1b with the interaction. The syntax is then: $WLI_{(per\ epoch\ bundle;\ N\ of\ data\ points\ per\ participant\ vary;\ standardized)} \sim 1 + condition_{(structured/random)} * epoch\ bundle_{(1:N\ of\ epochs)} + (1 + condition \mid participant)$. If we reached our maximum sample size and ~~that~~ the model does not yield reliable results crossing a threshold for H1/H0, or does not	(5) Our updating approach will yield multiple BFs. If the BF has not reached a threshold value at maximum sample size, we could possibly see an increasing or decreasing trend in the BF that provides more information than one BF for one sample size alone.	but not the random condition. Evidence for n~~N~~o interaction would indicate that, contrary to previous research, we found no evidence of progressive learning. This could mean that participants are at ceiling level of learning early on, or that they did not ~~learn the words~~become sensitive to the word structures at all (which should then be indicated by evidence for a null effect for condition; H1a).	

			converge, we will remove the interaction and test for the condition effect (H1a) alone. ~~To calculate the BF, we will follow Silvey et al. (2022). We specify our model of H1 as a half-normal distribution with a mean of 0 and an SD of 0.19 / 2 = 0.095, corresponding to the estimate for the original condition effect of Batterink & Paller (2017).~~ For the interaction effect, we will follow the same procedure for calculating the BF as above, while our SD is 0.07~~ / 2 = 0.035~~. Our sensitivity analysis will also be the same, but the range we will try will be from 0 to 0.14, as that is twice as big of an effect as Batterink and Paller (2017). See the simulation supplement and **RQ1** below for the models yielding these estimates on the data			

			of Batterink & Paller (2017). See also section 2.4.1. in the report.			
RQ2. Do we find behavioral evidence of SL in our structured condition?	We hypothesize that our participants will show behavioral evidence of word segmentation in the structured condition. This would be indicated by two predicted results: **H2a:** Familiarity ratings being higher for words than part-words and subsequently non-words in the Rating Task. **H2b:** A RT facilitation effect towards the word-final syllable in the Target Detection Task (TDT).		We will have two behavioral tasks of SL outcomes: the rating task and the Target Detection Task (TDT). *Rating task*: We will test whether words were judged as more familiar than part-words and subsequently non-words by using a CLMM with the following syntax: Rating ~ word category + (1|participant) + (1|item) Because the rating task has not been analyzed with a CLMM before, we will use Bain, which makes use of the approximate adjusted fractional BF (Gu et al., 2021; Hoijtink et al., 2019). By taking the foils as intercept, we formulate the following informative hypothesis for Bain:	The sample size rationale for RQ2 is based on the following arguments: (1) The rating task and TDT have been much used in earlier research (e.g., Batterink & Paller (2017), N = 45; Moreau et al. (2022); N = 24) finding ~~significant~~ evidence of learning repeatedly. (2) In our behavioral pilot (N = 19) (appendix B) we found ~~significant~~ evidence of learning with the TDT task and a 2AFC task for our stimuli. (3) In our student pilot (N = 15), we found ~~significant~~ evidence of SL for our stimuli in both the rating task and the TDT. See the link to the student project at the top page above this table.	If we find the results hypothesized for the behavioral tasks, we interpret this as our participants acquiring the word structures through SL and showing behavioral evidence of learning. If we find no behavioral evidence of learning or unexpected patterns of learning, our interpretation will largely depend on the neural measurements of SL (RQ1). If we do find neural evidence of learning, we cannot say that no learning occurred, but perhaps that learning ~~was very implicit~~was insufficient to influence behavior.	Participants can acquire words through SL without instruction and behaviorally show indications of learning.

			H1: $\beta_{part\text{-}word} > 0$ & $\beta_{word} > 0$ & $\beta_{word} > \beta_{part\text{-}word}$. See section 2.5.1. for information on how Bain calculates the prior and posterior distributions, and the BF. After the initial analysis, we will also conduct 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 × *b* (default), to 2 × *b*, as well as 3 × *b*. If the BF does not substantially change, we can conclude that the results are robust (Hoijtink et al., 2019, pp. 548-549). See also section 2.5.1.~~After the initial analysis, we will also conduct a sensitivity analysis with the fractions 1, 2, and 3 (Hoijtink et al., 2019).~~ *TDT:* Only RTs ≤ 1200 ms after target onset will be considered for analyses.			

			Reaction times will be analyzed using a LMM with the following syntax: RT ~ within-word syllable $\text{position}_{(\text{initial, medial, final})}$ + syllable position in $\text{stream}_{(\text{trial number})}$ + (1|participant) We will use the same methodology for calculating the BF as in RQ1, 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). We will follow this analysis with a sensitivity analysis reporting a robustness region (Dienes, 2019). We will test for prior models of H1 where the RT difference is 0 to 150 ms to find the region where the BF is still > 3 or < 1/3. See section 2.5.1.			

RQ3. Is behavioral SL performance correlated with the WLI in the structured condition?	**H3a:** For the rating task, this is explorative. Earlier research did not find a ~~significant~~ conclusive correlation with the WLI (Batterink & Paller, 2017; Moreau et al., 2022). However, see the sample size justification for power calculations stating that we might be able to find a conclusive result. We do expect the correlation to be positive if it exists. **H3b:** For the target detection task, we do hypothesize a positive correlation with the WLI, also in line with earlier research (Batterink & Paller, 2017; 2019).		We will conduct correlation analyses between the overall WLI in the structured condition, the rating score, and the RT facilitation score to determine whether individual variability in neural entrainment during exposure is related to subsequent behavioral SL performance. ~~For the correlation analyses,~~ We will test for a positive correlation in JASP. Since the effects in Batterink and Paller (2017) were $r = 0.32$ for the rating task, and $r = 0.42$ for the TDT, we will adhere to the prior κ = 0.5, which places less prior weight on big effect sizes and relatively more around 0. ~~we will adhere to the uniform default prior κ = 1.~~ We will follow this analysis with a sensitivity analysis that JASP provides. It calculates the BF over the range of possible prior values and plots these	We conducted a Bayes Factor Design Analysis (BFDA; Schönbrodt & Wagenmakers, 2018; Schönbrodt & Stefan, 2019) for the correlations reported in Batterink & Paller (2017) with regard to the rating score and RT facilitation score with the structured WLI. To find evidence for the reported correlations from Batterink and Paller (2017), we would need: - Rating score; $r = .32$, κ = 0.5, the Average Sample Number (ASN) at stopping point was N = ~~73~~66, $BF_{10} > 6$ in ~~94~~86.~~9~~5% of simulations. - TDT; $r = .42$, κ = 0.5, ASN = ~~55~~52, $BF_{10} > 6$ in ~~99~~98.8% of simulations. - H0, κ = 0.5: ASN = 82, $BF_{10} < 1/6$ in 55.2% of simulations, or ASN = 60, BF_{10} <	If we do not observe a ~~significant~~ correlation between the WLI and the rating task (e.g. an inconclusive BF or evidence for H0 that there is no correlation ~~but~~) but do find a ~~significant~~ positive correlation between the WLI and the target detection task, ~~it~~ this aligns with prior research. ~~This~~ If we find evidence for the absence of a correlation between the rating task and WLI, this suggests that the rating task involves explicit memory outcomes in SL, contrasting with the implicit nature of the target detection task and the WLI. Implicit learning appears linked to the neural measure of SL. Conversely, if we do find ~~significant~~ evidence for positive correlations between the WLI and both the rating task and the target detection task, it ~~implies~~ suggests that the WLI at least partially ~~can detect~~ reflects explicit	The WLI is related to (implicit) behavioral SL performance.

			results. See section 2.5.2 for more details, and the JASP correlation supplement for examples of this.~~We will follow this analysis with a sensitivity analysis.See section 2.5.2.~~	1/3 in 84.7% of simulations See also the **BFDA for correlations** clarification below this table.	memory of acquired words, alongside its sensitivity to implicit memory. In cases where there is evidence for no ~~significant~~ correlation between the WLI and the TDT, we will explore the WLI's time course further. If we confirm H1 and H2, this does not necessarily indicate no learning. Behavioral SL task performance reflects SL ~~abilities~~abilities but includes ~~as well as~~ other factors like meta-cognitive decision-making and memory retrieval.	
RQ4. Are our measures of individual differences correlated?	This is partially explorative. **H4:** We do hypothesize ~~significant~~ positive correlations between all tests for rhythmic ability: CA-BAT, PROMS, and SSS task.		We will perform a correlation analysis between all tests for individual differences: CA-BAT, PROMS, SSS task, Gold-MSI, Digit Span, and PPVT-III. For the correlation analysi~~e~~s between the rhythm tasks (H4), we will adhere to the ~~uniform default~~ prior κ = ~~1~~0.75,	See the **BFDA for correlations** clarification below this table for simulation-based power analysis for ~~a small,~~ medium, and large effect sizes. In our preliminary analysis of a student pilot sample (N = 15; see JASP supplement), substantial effect sizes, particularly in	If we find one or more ~~significant~~ positive correlations between the rhythm tests this indicates that they (to a large extent) measure the same individual capabilities. If we ~~do not~~ find negative correlations or evidence for H0 indicating the absence of ~~significant~~ correlations between the rhythm tests,	The tasks for individual differences, specifically the tasks for musical rhythm, are correlated.

	Whether there is a ~~significant~~ correlation between the rhythm tasks and the PPVT, Digit Span, and self-report questionnaire Gold-MSI is explorative, as well as correlations between these tasks themselves. If there are correlations, we expect them to be positive.		because we expect large effect sizes. For the correlations with the other tasks, we will adhere to the prior of κ = 0.5 because we do not expect those effects to be as large. We will follow these analyses with a sensitivity analysis that JASP provides. See RQ3 and section 2.5.2 for details. ~~We will follow this analysis with a sensitivity analysis.~~ We will investigate the contributions of each task to the WLI in the mediation analysis for RQ5. Only the tasks that show evidence that they correlate positively ~~significantly~~ with the SSS task will be used for that analysis. See section 2.6.	the rhythm-related tasks, were observed. While these findings are from a pilot study and should not be heavily relied upon, they suggest potential power for uncovering correlations, particularly among the rhythm tests (H4), which are part of our critical analyses.	this might indicate that these tests do not measure rhythmic ability in the same way. Perhaps other tasks used in this experiment are also inter-correlated.	
RQ5. Is rhythm perception related to SL	**H5a:** We hypothesize a direct effect of the		We will perform a mediation analysis in Lavaan (Rosseel, 2012) and Bain (Gu et al., 2021;	We calculated power for t he direct effect in multiple ways and follow the total effect heuristic from Dien	If there is a positive correlation between the PLV and the WLI in the structured condition, it	Rhythmic abilities ~~indicate~~ correlate with

performance as indicated by the WLI?	SSS task PLV on the WLI. **H5b:** We hypothesize that rhythmic and musical abilities have a positive influence on SL performance as measured with the WLI. We hypothesize that this is indicated by a direct effect of SSS PLV, mediated by rhythmic ability as measured by the CA-BAT, PROMS, and possibly also other tasks of individual differences if they correlate with the SSS task. See RQ4 for the selection procedure of possible mediators.		Hoijtink et al., 2019) with the WLI in the structured condition as the dependent variable. ~~We will test for a direct effect (*c-path*) of the SSS PLV~~ We will test for this direct effect initially by performing a regression of the SSS task on the WLI, and subsequently loading the model in the package Bain, under the informative hypothesis for the direct effect: *c-path* > 0. The hypothesis for a null effect will be defined as *c-path* = 0.~~first by calculating the Pearson's correlation coefficient with the WLI, using the default prior κ = 1. This is more conservative than using the pilot results as a prior (see RQ4 for discussion of the pilot effect sizes). We will follow this analysis with a sensitivity analysis.~~ If this yields a BF_{10} > 6, we will perform the full mediation analysis in Bain~~, under the informative~~	es (2019), which states tha t "Mathematically, the tota l effect is the sum of the di rect effect and the indirect effect. Thus, one possible theory is that the total effe ct is the maximum that co uld be expected for the dir ect effect." (p. 373). See the **Power Simulatio ns for Mediation** section below this table.	replicates and extends Assaneo et al.'s (2019) findings that 'high synchronizers' exhibit enhanced SL compared to 'low synchronizers.' The absence of this correlation could cast doubt on Assaneo et al. (2019)'s results or suggest design discrepancies, possibly due to our different stimuli and measurements. If we uncover indirect effects where rhythmic ability leads to a higher structured WLI, we interpret this as rhythmic ability positively influencing SL. If such effects are not present, depending on the direct effect of the SSS task, we can conclude either that speech synchronization is a superior predictor of SL compared to rhythm perception, or, that individual SL performance is better explained by rhythmic abilities than the SSS task.	~~better~~ SL ability.

			~~hypothesis for the direct effect: *c-path* > 0~~. We will add the tasks that correlated ~~significantly~~ positively with a BF_{10} > 6 with the SSS task in RQ4 as mediators. We will evaluate ~~them~~ the mediation in Bain under the informative hypotheses *a-path* > 0 & *b-path* > 0. After the initial analysis, we will also conduct a sensitivity analysis ~~with the fractions 1, 2, and 3 (Hoijtink et al., 2019)~~as described in RQ2 and section 2.5.1. See section 2.6.			
RQ6. Is working memory related to SL ability?	**H6:** We exploratively hypothesize that a larger working memory is related to a higher WLI in the structured condition.		We will perform a Bayesian correlation analysis between the WLI in the structured condition and the Digit Span if it is not included in the mediation analysis~~using the default prior κ = 1~~. For this correlation, we will adhere to the prior κ = 0.5, which places less prior weight on big effect sizes and relatively more around 0.	See **BFDA for Correlations** below this table.	If we find a ~~significant~~ positive correlation between the WLI and the digit span, we interpret that as working memory being a source of individual variability in SL. Conversely, a negative correlation or null effect, could be interpreted as an interference effect of working memory for SL, as	A larger working memory ~~indicates~~ predicts better SL ability.

			This gives us a reasonable chance of finding a theoretically interesting medium-to-large effect size if it exists (see also our simulations supplement and appendix A). We will follow this analysis with a sensitivity analysis that JASP provides. See RQ3 and section 2.5.2 for details. If the Digit Span is correlated with the tests included in the mediation analysis (RQ5), we will instead include it as a mediator. See section 2.6.		some previous research has found that depleted working memory can aid SL (see section 1.4).	
RQ7. Is better SL in adulthood related to having a larger vocabulary?	**H7.** This is explorative. In children, SL has been found indicative of vocabulary size. We want to test whether this also holds in adulthood.		We will perform a Bayesian correlation analysis between the WLI in the structured condition and the PPVT-III if it is not included in the mediation analysis ~~using the default prior κ = 1 using the default prior κ = 1~~. For this correlation, we will adhere	See **BFDA for Correlations** below this table.	If we find evidence for a ~~significant~~ positive correlation between the WLI and the PPVT-III, we interpret that as a positive relationship between SL ability and vocabulary size. A negative correlation or evidence for a null effect	SL ability relates to ~~is indicative of~~ vocabulary size, even in adulthood.

			to the prior κ = 0.5, which places less prior weight on big effect sizes and relatively more around 0. This gives us a reasonable chance of finding a theoretically interesting medium-to-large effect size if it exists (see also our simulations supplement and appendix A). We will follow this analysis with a sensitivity analysis that JASP provides. See RQ3 and section 2.5.2 for details. If PPVT-III is correlated with the tests included in the mediation analysis (RQ5), we will instead include it as a mediator. See section 2.6.		could be interpreted as an interference effect of the adult vocabulary for SL.	

Bayesian Updating: We chose 15 participants as the updating sample size, because this reflects approximately two to three weeks of data collection. Then, we use a third or fourth week to re-do the analyses and to determine if we need to add another sample. This way, we can create a monthly updating cycle. Our critical analyses determine the termination of data collection when they all reach a threshold BF_{10} of > 6 or ~~BF_{01}~~ < 1/6. We will collect data until this is the case for all these analyses, or until we reached a maximally feasible sample of 105 participants (45 + 4 updating cycles). These analyses (marked green in the table) are the following:

- The analyses for RQ1a, replicating the condition effect of Batterink & Paller (2017).
- RQ4, H4; correlations between the direct tests for rhythm perception; PROMS, CA-BAT and SSS~~, and possibly also the Gold-MSI~~.
- RQ5; a direct effect of SSS PLV on the WLI, so we are able to perform the mediation analyses for investigating the influence of rhythm perception on SL.

~~o RQ6 and RQ7; correlations calculated for the WLI with vocabulary and working memory if they are not added to the mediation.~~

RQ1 Sample Size Simulations
See the supplementary materials for the full R Markdown.
Condition effect:
- N = ~~20~~45; BF_{10} > 6 in 99~~2~~.~~9~~8% of simulations
- This increased to 100% for N = ~~50~~ 75 or more.

~~Interaction effect:~~
~~- N = 50; BF_{10} > 6 in 49.1% of simulations.~~
~~- N = 100; BF_{10} > 6 in 80.9% of simulations.~~
~~- N = 150; BF_{10} > 6 in 91.9% of simulations.~~

With regard to finding evidence for H0, this is always more difficult, especially with such a robust result from earlier research.
Condition effect:
- N = ~~50~~45; $BF_{\cancel{0}10}$ < 1/6 in ~~52~~48.~~6~~7% of simulations.
- For N = 105~~0~~ this became ~~68~~66.~~9~~2%.
~~- Finally, it did not increase much for N = 150, which was a BF_{01} < 1/6 in 73.3% of simulations.~~

~~Interaction:~~
~~- Even N = 150 yielded BF_{01} < 1/6 in only 26.4% of simulations~~

We argue that $BF_{\cancel{0}10}$ < 1/3 is moderate evidence for H0 and could also be a reasonable threshold for evidence if the maximum sample size is reached.
Condition effect:
- N = ~~100~~ 45 yielded $BF_{\cancel{0}10}$ < 1/3 in ~~86~~75.~~1~~4% and
- N = 105~~0~~ in ~~88~~85.~~5~~6% of simulations.

~~Interaction:~~
~~- N = 100; BF_{01} < 1/3 in 54.9% of simulations.~~
~~- N = 150; BF_{01} < 1/3 in 61.1% of simulations.~~

BFDA for correlations:

Small effect sizes seem unfeasible to detect in this project (see the simulations supplement). Such small effect sizes (r = .1) are also not meaningful as the critical analyses consist of the regression of the SSS task on the WLI and correlations between rhythm tasks, in order to do the mediation analysis. Small effect sizes in the order of .1 are not theoretically meaningful enough to become part of the mediation analysis.

Medium effect size r = 0.3 κ = 0.5:~~For a null correlation, we found that:~~
- ~~84~~81.~~3~~5% of simulations were stopped at $BF_{\cancel{0}10}$ ~~<~~> ~~1/~~6. ASN = ~~79~~70.
- H0 κ = 0.5: 55.2% of simulations were stopped at BF_{10} < 1/6. ASN = 82, or 84.7% of simulations were stopped at BF_{10} < 1/3. ASN = 60

Large effect size r = 0.5, κ = 0.75:
- 100% of simulations were stopped at BF_{10} > 6. ASN = 47.
- H0 κ = 0.75: 67.9% of simulations were stopped at BF_{10} < 1/6. ASN = 75, or 88.9% of simulations were stopped at BF_{10} < 1/3. ASN = 56
- ~~Adhering to BF_{01} < 1/3 yielded ASN = 56, and evidence for H0 in 93.3% of simulations.~~
- ~~For a very small effect size of r = .1, we will be unable to find sufficient evidence, even if our maximum sample size would have been N = 150. Only 16.9% of simulations terminate at BF_{10} > 6, while 44.7% of simulations falsely provided evidence for H0 with a BF_{01} < 1/6. Therefore, we added r = .2 as a second small sample size for simulations. This yielded ASN = 103, with 54.6% of simulations showing BF_{10} > 6.~~
- ~~For a moderate effect size of r = .3, ASN = 78, BF_{10} > 6 in 91.9% of simulations.~~
- ~~For a large effect size r = .5, ASN = 48, BF_{10} > 6 in 100% of simulations.~~

Power Simulations for Mediation

(1) Assaneo et al. (2019, p. 4) reported an effect size of r = .4 for the rank-biserial correlation comparing high-synchronizers with low-synchronizers for SL performance on a 2AFC task. This underestimates the effect we will investigate: SSS PLV and WLI, as these are more direct measures of SSS and SL. Yet, we performed BFDA on this effect size with κ = 0.5 and found ASN = 54~~7~~, with BF_{10} > 6 in 97~~9~~.~~6~~9% of simulations.

(2) We ran a linear regression WLI ~ SSS PLV on the student pilot data (N =15). This yielded a ~~significant~~ positive effect of SSS PLV, with an estimate of 0.63. ($R2$ = .40, F(1, 13) = 8.78, p = .011). We simulated data with a similar correlation of +/- .63, loaded it in Bain and found:

- BF_{10} > 6 in ~~99.5~~100% of simulations from N = ~~50~~ 45 onward for the hypothesis $\beta_{SSS} > 0$ ~~increasing to 100%~~.
- For H0 with the hypothesis $\beta_{SSS} = 0$ we found $BF_{\cancel{0}10}$ ~~< 1/6~~> 6 for N = ~~100~~ 45 in ~~64~~35.~~6~~1%, and N = 105~~0~~ in ~~69~~70.~~8~~9% of simulations.
  - For $BF_{\cancel{0}10}$ ~~<~~> ~~1/~~3 this was N = ~~100~~ 45 in ~~82~~79.~~6~~3% and N = 105~~0~~ in ~~85~~87.~~2~~8% of simulations.

(3) The estimate of 0.63 is identical to the correlation of the WLI and SSS PLV. Therefore, we used BFDA again, with a prior of κ = 0.75: For r = .63, ASN = 46, BF_{10} > 6 in 100% of simulations. H0, κ = 0.75: ASN = 75, BF_{10} < 1/6 in 67.9% of simulations.

(4) See **BFDA for Correlations** above for the BFDA with ~~zero, small,~~ medium, and large effect sizes.

Appendix B. Pilot Study

We conducted a behavioral pilot study using the same stimuli as the proposed experiment and other stimuli suitable for SL word segmentation experiments (for our preregistration of this pilot, see van der Wulp et al., 2022). We performed a speech segmentation SL experiment, including the Dutch version of the Gold-MSI (Bouwer et al., 2016; Müllensiefen et al., 2014). The aims of this pilot study were (1) to confirm that we could observe expected SL effects at the behavioral level using our newly created stimuli adhering to Dutch phonotactics, (2) to assess whether there were significant differences in SL between stimulus versions, and (3) to test for a possible first behavioral indication that musical sophistication is associated with better SL underlying word segmentation.

B.1. Pilot participants

A total of 19 participants took part in the pilot study, of which 14 were female, 4 male and 1 participant did not wish to specify their gender. None of the participants reported having AD(H)D, dyslexia, or other concentration- or language-related problems. All participants were native speakers of Dutch and over 18 years old (*M* = 25.6; *SD* = 9.8). The pilot experiment was approved by the Linguistics Chamber of the Faculty Ethics Assessment Committee of Humanities at Utrecht University (reference number: 22-031-03), and participants were compensated with €5 for their time (30 minutes).

B.2. Pilot stimuli

The stimuli used in the pilot study are identical to the stimuli as described in section 2.2, in the main manuscript, except that the pilot contained more versions of these stimuli. The syllable inventories were named *A* and *B*, with each three versions of words in the structured condition, of which only *A.1* is proposed to be used in the structured condition of the main experiment, and the syllables set *B* are used in the random condition, by randomizing their order of presentation (see section 2.2). See Appendix C for both syllable inventories and the words used in the Pilot study. For more information on the creation of these stimuli, we refer to the preregistration for this experiment (van der Wulp et al., 2022). In the pilot experiment, each inventory had three structured versions to be counterbalanced between participants in order to prevent effects of syllable idiosyncrasies. However, as we aim to investigate individual differences in this experiment, we decided to choose one version of the stimuli for the structured condition. In the pilot experiment, we thus had six stimulus versions, which were counterbalanced over our 19 participants such that three participants listened to each version, with the exception that four participants listened to version *B.3*. The methodology for creating the audio files was the same as described in section 2.2.

B.3. Pilot procedure

B.3.1. *Listening task*

Each participant listened to one of the stimulus versions described in section B.2 and Table C3. Transitional probabilities for syllables within words were 1.0 and between words 0.33. The words were presented in a pseudorandom order, so the same word did not repeat consecutively. The listening task was divided into three blocks of four and a half minutes. Between these blocks, participants took untimed breaks.

B.3.2. *Two-alternative forced choice task*

After the listening task, participants performed a two-alternative forced choice task (2AFC task), which is assumed to gauge participants' explicit memory of the words in the stimuli. In each trial, participants chose one out of two words presented auditorily: one being a trisyllabic word from the listening task and the other being a part-word or non-word foil created with the syllables from the same inventory. Subsequently, participants were asked to rate on a four-point scale how familiar the word they chose was to them. The stimuli used in this 2AFC task are shown in table C4 in Appendix C. There were two part-words, two non-words, and then the four words that were presented during the listening task. These words and foils were combined exhaustively into 16 trials. We predicted that our participants would show a significant preference for words, compared to part-words and non-words in the 2AFC task, along with an average accuracy significantly above chance (50% correct) indicating successful statistical word learning.

B.3.3. *Target Detection Task*

The second post-learning task our participants performed was a target detection task, almost identical to the task described in section 2.3.2. The target detection task in the pilot experiment was shorter. For each target syllable there were two speech streams, with the target occurring four times per stream, resulting in 24 speech streams and 96 targets for this task. We predicted that the reaction times of our participants would follow this pattern of facilitation towards the word-final syllable as a second behavioral indication of successful SL.

B.3.4. *Questionnaire and musical sophistication*

Participants filled out a questionnaire about their experience during the experiment and the stimuli they were presented with ('did it contain existing Dutch words?'), their (linguistic) background (native language, other languages mastered, age, educational level), and their musical sophistication, as measured with the Dutch translation of Gold-MSI (Bouwer et al., 2016; Müllensiefen et al., 2014).

B.4. Pilot data analysis

With respect to the 2AFC task, accuracy was computed for each trial based on the participant's choice for a word (accuracy = 1) or a non-word/part-word foil (accuracy = 0). This was summarized as percentages correct. We used a t-test to determine if our participants' performance was above chance level (50% correct). We also analyzed the results on the 2AFC task with a Binomial Logistic Regression using a Generalized Linear Mixed-Effects Model (GLMER) with the raw accuracy scores to assess the possible influence of stimulus version, foil type (non-word or part-word), and musical sophistication scores from the Gold-MSI on the 2AFC task accuracy, including a random intercept of participant. For the iterative model process, see table C5 in Appendix C.

For the target detection task, reaction times (RTs) were calculated for each participant and target syllables with respect to detected targets ("hits;" within 0-1200 ms after target onset following Batterink & Paller, 2019) in each within-word syllable position (word-initial, word-medial, and word-final). The target detection task was analyzed using a Linear Mixed Model with RT as the dependent variable and within-word syllable position (word-initial, word-medial, and word-final) as the predicting factor, to establish whether the facilitating effect of the word-final syllable position – indicating statistical learning – is present in our data. We included random intercepts of participant and syllable. As a subsequent step, we added stimulus version and the Gold-MSI scores as predictors as well. For the iterative model process, see table C6 in Appendix C.

B.5. Pilot results

B.5.1. *2AFC task results*

With respect to the 2AFC task, our participants scored on average 62.5% correct (SD = 17.3%). This is significantly above chance (t (18) = 3.15, p = .006). However, a Shapiro-Wilk test indicated that our data was not normally distributed (W = 0.90, p = 0.05). This was due to one outlier participant as detected using the boxplot method[1]. Figure 4 shows the performance on the 2AFC task per participant. Without the outlier participant in our analysis, the average score on the 2AFC task increased to 65.3% correct (*SD* = 12.72%), again significantly above chance (t (17) = 5.09, p < .001). Next, we wanted to investigate if the stimulus version influenced the accuracy scores. Participants exposed to syllable inventory A scored 63.89% (*SD* = 1.96%) correct, and participants exposed to syllable inventory B scored 66.67% (*SD* = 1.96%) correct. These averages are not significantly different from one another (t (15.70) = -0.45, p = .66). We further checked this for all sub-versions (*A.1, A.2, A.3, B.1, B.2, B.3*) by performing a one-way ANOVA to compare the effect of stimulus version on the 2AFC scores, both with and without the outlier participant. Both one-way ANOVAs revealed that there was no statistically significant difference in scores between the groups (without outlier: F (5, 12) = 1.76, p = .20; with outlier: F (5, 13) = 1.63, p = .22).

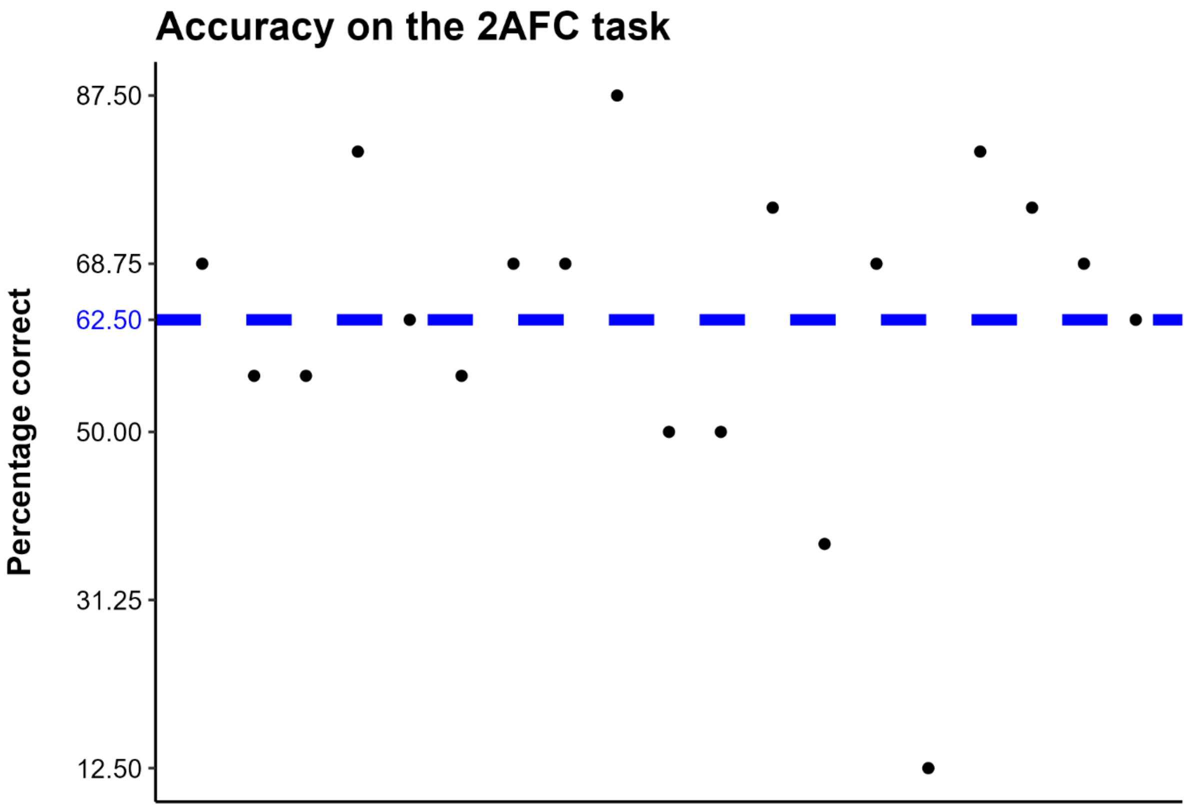


Figure 4. Performance on the 2AFC task. Each dot represents one participant. The blue line is the average percentage correct. This plot includes the outlier participant (12.5% correct).

[1] In the boxplot method, values above Q3 + 1.5xIQR or below Q1 - 1.5xIQR are considered to be outliers.

For our LMM analyses, we iteratively added predictors and used the likelihood ratio test of the model's fit to the data to determine if an added factor improved the model ($p \leq .02$; Winter, 2020). The model iterations can be viewed in table S5 in the Supplementary Materials. Our final model for the 2AFC task data was a Binomial Logistic Regression GLMER which had accuracy as the dependent variable and a random intercept for participant.[2] Foil type was a significant predictor (*OR* = 0.49, 95% CI [0.30, 0.80], $z = -2.84$, $p = .004$), indicating that in our sample, part-words were more difficult than foils to correctly reject.

B.5.2. *Target detection task results*

The average reaction times (RTs) of our participants showed the expected pattern of facilitation towards the word-final syllables (see Figure 5). We statistically confirmed this with a Linear Mixed Model (LMM) having RT as the dependent variable. The iterations for this model can be viewed in the Supplementary Materials; table S6. The final model included within-word

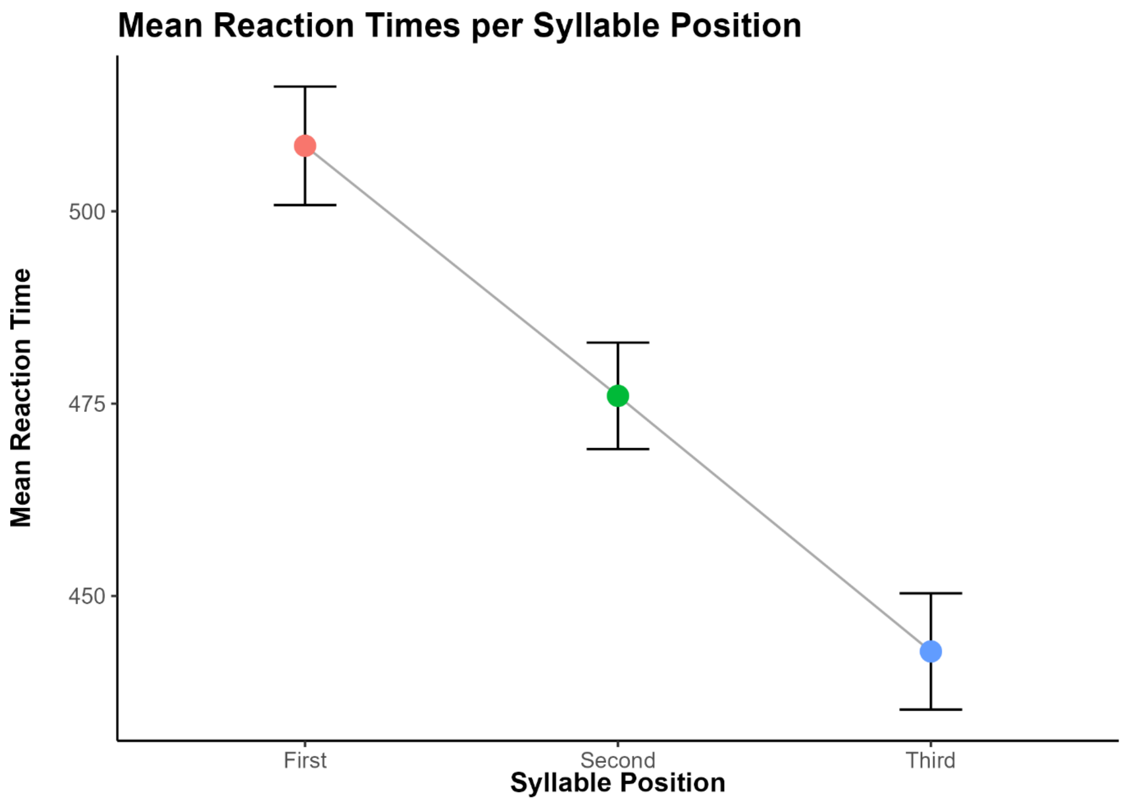


Figure 5. Average RTs per syllable position on the target detection task. The error bars reflect the Standard Error of the Mean (SEM).

syllable position as a predictor and random intercepts for both participant and target syllable.[3] The effect of within-word syllable position was highly significant, indicating the expected

[2] Formula of the final 2AFC model: accuracy ~ foil type + (1|participant)
[3] Formula of the final RT model: RT ~ syllable position + (1|participant) + (1|syllable)

facilitation towards the word-final syllable (see Figure 5; $\beta = -31.91$, $t(1528.29) = -6.60$, $p < .001$, 95% CI [-41.40, -22.43]). We again found no effects of stimulus version (*A* & *B* but also within these sets versions 1-3: $p > .05$, see table C6).

B.5.3. *Musical sophistication results*

Our pilot participants were not highly musical and none of them was a professional musician. This was reflected by a mean score of 3.15 ($SD = 0.96$) on the Gold-MSI, which ranged between 1.56 and 5.05. The possible range of scores on the Gold-MSI is between 1 and 7. We centered the general score on the Gold-MSI and added it as a predictor to the models of the 2AFC task and target detection task, but this did not significantly improve these models (see Table S5 and S6).

B.6. Pilot discussion and conclusion

The results of our pilot experiment indicate that participants successfully acquired the word forms presented during the listening task. Behaviorally, they demonstrated that they could accurately discriminate the words from part-word and non-word foils during the 2AFC task. It was more difficult for the participants to correctly reject part-words than non-words. This is expected as the part-words were present in the stimulus streams, but do not allow segmentation according to the transitional probabilities of the input. Furthermore, the target detection task fully followed our predicted pattern of facilitation towards the final syllable of the word. In addition, there was no evidence of significant differences between stimulus versions in neither the 2AFC task nor the target detection task (see tables S5 and S6).

The Gold-MSI was not a significant predictor for any of the tasks. However, it approached significance for the target detection task ($p = .06$; table S6). We therefore hypothesize that target detection performance – as an indication of implicit memory of word forms as expressed by a facilitation pattern towards the final syllable of the word – will be significantly enhanced (e.g., facilitation will be steeper) in musically trained individuals if more participants are included and thus statistical power is increased. This was beyond the scope of this pilot experiment, but the experiment proposed in section 2 will of course investigate this further and combine it with an online measure of SL using EEG, which might be more sensitive to an influence of musicality than the offline RT task used in the pilot. Moreover, we will use two more specific rhythm processing tasks in this follow-up experiment as well, which we hypothesize will be more directly related to SL performance than a self-report general measure of musicality.

In summary, the pilot experiment indicates that our new stimulus set – adhering to Dutch phonotactics – is suitable for SL word segmentation experiments with native Dutch speakers. Behavioral performance in both explicit and implicit memory tasks indicated that our participants were able to acquire the words based on transitional probabilities in the absence of other phonological cues such as intonation or pauses.

Appendix C

Table C1a.

Set A and EEG markers.

Syllable	ID
ba	10
bo	11
by	12
χø	13
χi	14
mø	15
ta	16
tø	17
ti	18
to	19
sy	20
su	21

Table C1b.

Set B and EEG markers.

Syllable	ID
da	22
dø	23
dy	24
χo	25
χy	26
nu	27
pø	28
py	29
ro	30
sa	31
sø	32
ri	33

Table C2

Items for the rating task

Item	Category
suχita	word
tobamø	word
sytøbo	word
χøbyti	word
tatoba	part-word foil
tøboχø	part-word foil
møsyχi	part-word foil
bytisy	part-word foil
χitato	part-word foil
bamøsu	part-word foil
boχøby	part-word foil
tisytø	part-word foil
tatøχø	non-word foil
boχito	non-word foil
møbysu	non-word foil
tibasy	non-word foil

Table C3.

Stimuli for the pilot experiment

Version	Syllable position		
	1	*2*	*3*
A.1.	su	χi	ta
	to	ba	mø
	sy	tø	bo
	χø	by	ti
A.2	ta	su	χi
	mø	to	ba
	bo	sy	tø
	ti	χø	by
A.3	χi	ta	su
	ba	mø	to
	tø	bo	sy
	by	ti	χø
B.1	da	pø	nu
	dø	χo	py
	ro	dy	sa
	χy	ri	sø
B.2	nu	da	pø
	py	dø	χo
	sa	ro	dy
	sø	χy	ri
B.3	pø	nu	da
	χo	py	dø
	dy	sa	ro
	ri	sø	χy

Table C4

Part-words and non-words used in the 2AFC task of the pilot experiment.

Language	Part-words	Non-words
A.1	tatoba	tømøsu
	tøboχø	tiχibo
A.2	χimøto	tømøsu
	sytøti	tiχibo
A.3	subamø	tømøsu
	bosyby	tiχibo
B.1	nudøχo	pøχydy
	dysaχy	riχonu
B.2	pøpydø	pøχydy
	rodysø	riχonu
B.3	daχopy	pøχydy
	sarori	riχonu

Table C5
Model summary pilot study 2AFC task of the pilot experiment

Nr.	-2LL	Nr. of Parameters	Model fit *p* (χ 2 dist.)	Model comparison	Predictor added	Action
Model 0	-198.39	2			Random intercept participant	keep
Model 1	-196.82	3	0.07	not better	Random intercept trial	remove
Model 2	-193.57	7	0.08	not better	Stimulus version	remove
Model 3	-194.75	3	0.007	better	Foil type	keep
Model 4	-193.62	4	0.13	not better	Gold-MSI score	remove

Table C6
Model summary pilot study target detection task of the pilot experiment

Nr.	-2LL	Nr. of Parameters	Model fit *p* (χ 2 dist.)	Model comparison	Predictor added	Action
Model 0	-10192	3			Random intercept participant	keep
Model 1	-10176	4	< .001	better	Random intercept syllable	keep
Model 2	-10154	5	< .001	better	Syllable position	keep
Model 3	-10149	10	0.06	not better	Stimulus version	remove
Model 4	-10152	6	0.06	not better	Gold-MSI score	remove